Latent Dependency Mining for Solving Regression Problems in Computer Vision

Ke CHEN

Submitted to the University of London in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Queen Mary University of London

2013

Latent Dependency Mining for Solving Regression Problems in Computer Vision

Ke CHEN

Abstract

Regression-based frameworks, learning the direct mapping between low-level imagery features and vector/scalar-formed continuous labels, have been widely exploited in computer vision, e.g. in crowd counting, age estimation and human pose estimation. In the last decade, many efforts have been dedicated by researchers in computer vision for better regression fitting. Nevertheless, solving these computer vision problems with regression frameworks remained a formidable challenge due to 1) feature variation and 2) imbalance and sparse data. On one hand, large feature variation can be caused by the changes of extrinsic conditions (i.e. images are taken under different lighting condition and viewing angles) and also intrinsic conditions (e.g. different aging process of different persons in age estimation and inter-object occlusion in crowd density estimation). On the other hand, imbalanced and sparse data distributions can also have an important effect on regression performance. Apparently, these two challenges existing in regression learning are related in the sense that the feature inconsistency problem is compounded by sparse and imbalanced training data and vice versa, and they need be tackled jointly in modelling and explicitly in representation. This thesis firstly mines an intermediary feature representation consisting of concatenating spatially localised feature for sharing the information from neighbouring localised cells in the frames. This thesis secondly introduces the cumulative attribute concept constructed for learning a regression model by exploiting the latent cumulative dependent nature of label space in regression, in the application of facial age and crowd density estimation. The thesis thirdly demonstrates the effectiveness of a discriminative structured-output regression framework to learn the inherent latent correlation between each element of output variables in the application of 2D human upper body pose estimation. The effectiveness of the proposed regression frameworks for crowd counting, age estimation, and human pose estimation is validated with public benchmarks.

Submitted to the University of London in partial fulfilment of the requirements for

the degree of Doctor of Philosophy

Queen Mary University of London

2013

Declaration

I hereby declare that this thesis has been composed by myself and that it describes my own work. It has not been submitted, either in the same or different form, to this or any other university for a degree. All verbatim extracts are distinguished by quotation marks, and all sources of information have been acknowledged.

Some parts of the work have previously been published as:

- Chapter 2
  - C. C. Loy, **K. Chen**, S. Gong and T. Xiang, Crowd Counting and Profiling: Methodology and Evaluation, in S. Ali, K. Nishino, D. Manocha, and M. Shah (Eds.), *Modeling, Simulation, and Visual Analysis of Large Crowds*, Springer, 2013, To Appear.
- Chapter 3
  - **K. Chen**, C. C. Loy, S. Gong and T. Xiang, Feature Mining for Localised Crowd Counting, in *British Machine Vision Conference* (BMVC), pp. 21.1-21.11, 2012.
- Chapter 4
  - **K. Chen**, S. Gong, T. Xiang and C. C. Loy, Cumulative Attribute Space for Age and Crowd Density Estimation, in *IEEE International Conference on Computer Vision and Pattern Recognition* (CVPR), 2013.
- Chapter 5
  - **K. Chen**, S. Gong and T. Xiang, Human Pose Estimation Using Structural Support Vector Machines, in *IEEE International Conference on Computer Vision, Workshop on Socially Intelligent Surveillance and Monitoring* (ICCV-SISM), pp. 846-851, 2011.

Acknowledgements

Firstly, I would like to thank my supervisors Prof. Shaogang Gong and Dr. Tao Xiang for their perpetual patience, encouragement and guidance. During the period of pursuing Ph.D degree, their invaluable time and efforts were spent on supervising me to do the research both technically and philosophically. Their invaluable advice and consistent support help me to construct my style of doing research in a logical and organised manner. I would like to express my deepest respect and many thanks from the bottom of my heart to them for their work.

In addition, I would like to thank Dr. Pengwei Hao for being my internal examiner throughout my Ph.D project with his encouragement and suggestion. Besides, many thanks are particularly given to my collaborator Dr. Chen Chang Loy for his efforts dedicated to my publications. My sincere appreciation goes to various seniors and juniors at Vision Group for their friendship and support, in particular Dr. Tim Hospedales, Dr. Wei-Shi Zheng, Dr. Somboon Hongeng, Dr. Jian Li, Dr. Tom Haines, Dr. Chris Russell, Dr. Anastasios Roussos, Dr. Sara Vicente, Dr. Lukasz Zalewski, Dr. Yogesh Raja, Dr. Samuel Pachoud, Dr. Khalid Bashir, Dr. Matteo Bregonzio, Dr. Bryan Prosser, Dr. Parthipan Siva, Yanwei Fu, Xiatian Zhu, Zhiyuan Shi, Yi Li, Xun Xu, Ryan Layne, Ravi Garg, Nikolaos Pitelis.

In addition, I sincerely and gratefully thank all friendly and highly competent administrative and systems support staff in the department for enabling things to run smoothly and efficiently.

I respectfully appreciate Prof. Yunong Zhang at Sun Yat-sen University for his inspiring encouragement to lead me to the path of academic career.

Finally, I would express my special thanks to my father Shoukun Chen and my mother Ying Li for their enduring love, support and understanding. Without them, I would never have completed my study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A number of computer vision problems, e.g. analysis of a crowd pattern in a scene, biometrics and human behaviours, can be solved by computer analysis of video/image data from cameras. These computer vision problems are still challenging and attract wide attention and great interests in sociology, psychology, and especially public safety. Specifically, analysis on crowd dynamic patterns aims to learn the crowd flow evolvement and floor fields [4], to track an individual in a crowd [142], to segment a crowd into semantic regions [112, 196], to detect salient regions in a crowd [113], or to recognise anomalous crowd patterns [90, 122]. For biometrics analysis, one aims to identify persons from the existing records [77], to verify the facial image of the person with another image [77, 91], or to estimate the age of persons to sell age-restricted products. For human behaviours understanding, one aims to estimate the human posture [45], to recognise the gait patterns of pedestrians [75], to detect the motion of humans [15], to track human motions in a video [62], or to re-identify the objects from disjoint cameras [96].

Among these computer vision problems, a common requirement is to estimate the scalar/vector-formed continuous values given low-level imagery features. In this thesis, we focus on three kinds of such problems: i.e. crowd density estimation (to count the number of persons in a scene), facial age estimation (to automatically estimate the ages of human from facial images), and human pose estimation (to automatically estimate the gesture of human body parts in still images). Solving these computer vision problems can be viewed as the preliminary basis of high-level recognition problems such as human behaviour recognition. Moreover, in our everyday life, these computer vision problems are highly related to public security and tragedies could occur

Figure 1.1: Example of surveillance footage frames captured during the Love Parade music festival in Germany, 2010, before the fatalities occurred.

Figure 1.2: Example of French parade against gay marriage/adoption in 2013.

when we overlook them.

1.1 Motivation

The motivation of this thesis will be presented in this section with focusing on three computer vision problems: i.e. crowd counting, age estimation and human pose estimation.

1.1.1 Crowd Counting

Tragedies involving large crowds often occur, especially during religious, political, and musical events [70]. For instance, a crowd crush at the 2010 Love Parade music festival in Germany, caused a death of 21 people and many more injured (see Figure 1.1). And more recently a stampede happened near the Sabarimala Temple, India with death toll crosses hundred. These tragedies could be avoided, if a safer site design took place and a more effective crowd control was enforced. Video imagery based crowd counting can be a highly beneficial tool for early detection of over-crowded situations to facilitate more effective crowd control. Moreover, the total number of the crowd in political or religious parade is also significant, which could change the support from the public, e.g. French parade against gay marriage/adoption in 2013 as shown in Figure 1.2. Police says there were around 340, 000 demonstrators while people who organised the parade performed a counting and found around 1, 000, 000 participants. However, counting the exact number of participants is extremely time-consuming and expensive by humans. This

(a) Age 2 (b) Age 14 (c) Age 28 (d) Age 40

Figure 1.3: Illustrative examples of FG-NET dataset for facial age estimation.

encourages the researchers to develop an effective and efficient way via surveillance videos and computer-aided technologies. It also helps in profiling the population movement over time and across spaces for establishing global situational awareness, developing long-term crowd management strategies, and designing evacuation routes of public spaces. In retail sectors, crowd counting can be an intelligence gathering tool [161] to provide valuable indications about the interest of customers through quantifying the number of individuals browsing a product, the queue lengths, or the percentage of a store's visitors at different times of the day. The information gathered can then be used to optimise the staffing need, floor plan, and product display.

1.1.2 Age Estimation

For age estimation given facial images, the problem is significant as age is one of the vital biometric information for human in numerous applications of computer vision such as facial verification [91] in attribute form. From the general public viewpoint, it is also necessary to develop an accurate and robust estimator for selling some age-restricted products (e.g. cigarettes and alcohol) by machines. However, compared to other biometric information such as fingerprint and iris with the identical representation for each person, facial images are ambiguous and do not have a unique evidence for a specific person. Intuitively, two persons could have similar appearance of faces but they do have different fingerprint and iris. As a consequence, a question might rise: why we need to exploit such an ambiguous cue with faces for age estimation? From common sense, faces contain more informative visual cues than fingerprint and iris, especially having shape and texture information, which can describe the aging progress well. In other words, fingerprint and iris do not have explicit sematic meaning associated with the aging process. As shown in Figure 1.3, the changes of appearance of the same person are illustrated. Evidently, accurate age

Figure 1.4: Examples of theft. Images from Internet.

estimation can also help to analyse the aging process pattern of different persons, which have its significance in physiology studies. Intuitively, the changes of facial bones and skull have evident relation to the status about getting mature.

1.1.3 Human Pose Estimation

Human pose estimation attracts public and researchers' attention owing to its significant value to surveillance applications. Specifically, the problem of estimating the configuration of human body parts accurately is a preliminary step and the basis for high-level recognition problems. Moreover, other computer vision problems such as tracking [99] and human detection [76] can also benefit from the success of addressing human pose estimation appropriately, because of its shared characteristics across the problems. Compared to the hardware of the visual surveillance in the public places, the manpower spent on monitoring and analysing the human is relatively more expensive. Previously, those surveillance cameras are only used for passively recording the data instead of detecting the events in real-time and taking action immediately [73,78,84,86,172]. For instance, if there happens a case of theft at the airport, currently we only refer back to the video records and find out the bio-characteristics of the theft. In Figure 1.4, some examples of theft are illustrated, which have similar posture. It is evident that such a passive application of those surveillance cameras is less efficient. Instead, if the surveillance systems react in real-time, the security staff would arrest the suspects in time, which save lots of manpower and further increase public safety. Consequently, human pose estimation gives the configuration of human body parts, which is the direct and most relevant evidence for crimes or other abnormal behaviours. Furthermore, body gestures can also give a lot of information besides the language. For examples, hand gestures are widely used for disabled people in the world [135].

1.2 Problem Definitions

The three aforementioned computer vision problems can be formulated into discriminative regression frameworks by learning a mapping function between a low-level imagery feature input vector and a continuous scalar/vector-formed output. The problems of video imagery based crowd counting in crowded scenes for population profiling, facial age estimation, and 2D human upper body pose estimation in still images remain non-trivial. Besides the characteristics of specific problems, the two common challenges shared by all three problems are feature variation and spare & imbalanced data.

Feature representations are ambiguous with large variation, which are caused by both the inconsistency of extrinsic conditions including lighting conditions and viewing angles and also intrinsic conditions of the specific problems as the following.

- For crowd density estimation, inter-object occlusion in a crowded scene and spatial density distribution (i.e. crowd density in different locations are different) can make the problem difficult.
- For age estimation, aging process of different persons is extremely different in addition to different hairstyle, glasses, gender and racial groups.
- For human gesture estimation, intrinsic conditions including occlusion between body parts and un-controllable environment can cause the large variation of feature representation.

For sparse and imbalanced data distribution, solving the problems of crowd counting, age estimation, and human pose estimation suffer from the existing datasets. The challenges are caused by either unreliable annotation or labourious ground truth annotating. Specifically, it is easy to find a large number of facial images from Internet, but annotating the truth age of each image is not quite reliable. For crowd counting and human pose estimation, label annotation is more reliable but labourious for either exhaustively head dot-annotation in the scene for crowd counting or annotating each body parts when multiple persons within the images. The difficulties in annotating the data cause a sparse and imbalanced data distribution in the existing datasets.

1.3 Contributions

These two challenges are related. Particularly, sparse and imbalanced data distribution can cause large feature variation, and vice versa. As a result, to tackle both challenges jointly in both model

learning and representation can help to improve the model performance. The contributions of this thesis are as follows.

Firstly, based on multivariate ridge regression, feature mining and information sharing from neighbouring localised regions are exploited for localised crowd counting. This work is achieved by considering a single multi-output ridge regression model for localised crowd counting which has advantages over both existing global approaches in providing local estimates and existing local approaches being more scalable. Moreover, with concatenated intermediary feature representation, the challenge of sparse and imbalanced data locally is mitigated in the sense that the localised regions without sufficient samples can seek support from neighbouring regions. More details are presented in Chapter 3.

Secondly, for the first time, an attribute representation is constructed for learning a regression model for facial age estimation as well as crowd density estimation. More specifically, a novel concept of cumulative attributes is proposed with both clear semantic meaning and also discriminative, with added advantages of efficiently computable and requiring no additional annotation. More importantly, such cumulative attribute space can cope with the challenges of both feature variation and sparse and imbalanced data jointly by capturing the cumulative dependent nature of label space in regression. More details are presented in Chapter 4.

Thirdly, for human upper-body gesture estimation, structural support vector machines are applied with much faster inference procedure than generative methods, which are more suitable for real-time applications. In addition, such a framework can generate acceptable results when the size of training database is reduced dramatically (i.e. make the data more sparse and imbalanced). Compared to non-structured discriminative methods (e.g. Support Vector Regression), structured methods achieve better performance owing to the ability to capture the important relevance information between outputs. More details are presented in Chapter 5.

1.4 Structure of the Thesis

In this thesis, we provide a comprehensive review of new regression-based frameworks for addressing the problems of localised crowd density estimation, facial age estimation, and 2D human posture estimation, and also give insights through extensive experiments.

Chapter 2 gives a structured critical overview of different approaches to three computer vision based applications as well as related regression techniques. Firstly, we present crowd count-

ing reported in the literature, including pedestrian detection, coherent motion clustering, and regression-based learning. In particular, we focus on the regression-based techniques that have gained considerable interest lately owing to their effectiveness in handling more crowded scenes. Secondly, for facial age estimation, both classification and regression based frameworks are investigated with more concern on regression frameworks. Thirdly, the existing techniques of human pose estimation are reviewed including body-parts detection-based and regression based learning. Finally, some widely-used regression techniques are also reviewed in this chapter.

For three specific computer vision problems, we have proposed and developed three regression based frameworks with extensive experiments conducted on public benchmarks. More details about these techniques and experiments are presented in Chapters 3-5. Conclusion and some remarks for possible future directions are given in Chapter 6.

Chapter 2

Literature Review

A number of computer vision problems can be addressed by a regression framework, i.e. learning the mapping between low-level/intermediate features and continuous label values directly. There are several key components for solving regression problems in computer vision as the following:

1 low-level feature extraction including background modelling, foreground highlighting and/or perspective normalisation;

2 intermediate representation if necessary, e.g. to contain location information;

3 continuous label representation;

4 appropriate regression techniques selection to solve the computer vision problems effectively and efficiently.

Considering different computer vision problems investigated in this thesis, we will give detailed review on the common characteristics and difference of three computer vision problems, which can thus be structured as the following. Firstly, Section 2.1 will give an introduction to background modelling and foreground highlighting to achieve more robust feature representation by removing the background and also reduce the search space of foreground. Imagery feature representation and label outputs adopted for solving the computer vision problems will be reviewed at the current stage of literature in Section 2.2. In Section 2.3, the review about perspective normailisation is given, which is significantly important in crowd density estimation. Section 2.4 will present the existing regression frameworks applied to crowd density estimation

problem as well as other state-of-the-arts in detection or clustering fashion. We then will investigate the regression and classification techniques used in facial age estimation in Section 2.5. The literature review of the-state-of-arts for human pose estimation will be given in Section 2.6. Finally, in Section 2.7, a review about widely-used regression techniques will be investigated. Before ending this chapter, the summary section will be concluded with several remarks.

2.1 Background Modelling and Foreground Detection

Background modelling and foreground highlighting is an important pre-processing step for solving a number of computer vision problems such as video analysis [19] and human pose estimation in still images [49]. In this section, we will give a brief introduction to the existing approaches of background substraction and foreground detection.

2.1.1 Background Subtraction

In video sequences, background subtraction is an efficient and effective way to capture the motion of objects with creating a binary foreground mask to segment foreground objects from background. Intuitively, each pixel of the image is considered as consistent in its colour during the time. Any changes of colours in the frames are assumed caused by motion of foreground objects. Apparently, the aforementioned assumption can hold when the illumination of scenes is consistent and it is not valid in other complex situation such as indoor scenes [30]. Moreover, the low contract between background and foreground and activity patterns by still objects for a long period make the background subtraction still challenging.

The simplest way of background subtraction is based on the difference between frames, i.e. the difference between current frame and the previous frame. Such a frame difference method for background subtraction has the advantages of simple implementation and efficient to compute under the assumption that the motion of foreground objects is continuous with an appropriate frame rate. Different form frame difference without a unique background model, temporal averaging method [69] has a background model generated by averaging the frames during a period of time under the assumption that the constant pixel in time domain is viewed as the background. In this way, the foreground objects can be obtained by subtracting the current frames and background model. The advantages of temporal averaging method, which is widely adopted [19, 22], are more robust than frame difference with considering a number of frames over a period of

time, but less sensitive to lighting condition changes. Consequently, other methods such as $\Sigma - \Delta$ background estimation [118], Mixture of Gaussian [154, 155], and mixture of dynamic textures-based method [21] is to segment the foreground objects with relaxing linear dependency on the difference between pixel. Benefiting from the distribution of multiple Gaussian for each pixel [31, 155], dynamic changes in the background can be learnt effectively which can cope with more complicated situations, such as crowd density estimation in indoor scene with lighting condition changing from time to time [30].

2.1.2 Foreground Highlighting

The aim of foreground highlighting is to find the foreground area, which is similar to background modelling but has some differences in the processing procedure. Generally, we hope to estimate the density map in the whole image area, which can be used to highlight the possible foreground area, e.g. saliency in object detection [26]. Within the foreground segments, high-level visual problems such as object detection, human pose estimation, and recognition can be more efficiently addressed owing to the reduced search/foreground space than those method without foreground segmentation. For human pose estimation in cluttered environments [49], the method by progressively reducing foreground space is adopted with object detection with a bounding box and graph cut, which can significantly improve the performance of human gesture estimation. Different from background modelling in Section 2.1.1, foreground highlighting has an intermediate-level and more informative representation to further analyse other computer vision problems instead of a binary foreground mask.

2.2 Feature Representation

Before exploiting regression frameworks for addressing specific computer vision problems, we will encounter the issue of extracting and representing features from images or video frames. It is known that feature representation in regression frameworks should be sufficiently discriminative to learn the boundary of observations. In view of different characteristics of three kinds of computer vision problems, we will categorise the feature representation into global and local level. Note that, in general, the features can be employed globally or locally according to the need, but we treat the feature representation containing location information as local features here. For global features, image features are extracted from the whole image space, which do not

(a) Original Example Frame (b) Foreground Segmentation (c) Edge

Figure 2.1: An illustrative example of original frame, foreground segment and edge from Mall dataset [30].

contain any spatial information. Generally, these global features are low-dimensional and thus inexpensive to learn the regression models, and suitable for addressing computer vision problems such as facial age estimation [51] and head pose estimation [66, 164]. For local features, the location information is incorporated with the price of computational cost. However, for some frameworks, location information can have an important effect on the performance, e.g. localised crowd density estimation [101]. On balance, the usage of global or local features are dependent on the specific computer vision problems and we will give a detailed review on some feature representation employed in the experiments of this thesis.

2.2.1 Foreground Segmentation Features

After pre-processing background modelling, foreground segments with an example shown in Figure 2.4(b) can be obtained. Various holistic features can be derived from the extracted foreground segments, for example, as:

- Area – total number of pixels in the segment.
- Perimeter – total number of pixels on the segment perimeter.
- Perimeter-area ratio – ratio between the segment perimeter and area, which measures the complexity of the segment shape.
- Perimeter edge orientation – orientation histogram of the segment perimeter.
- Blob count – the number of connected components with area larger than a predefined threshold, e.g. 20 pixels in size.

Various studies [19, 40, 116] have demonstrated encouraging results using the segmentation-based features despite its simplicity. Several considerations, however, has to be taken into account

during the implementation. Firstly, to reduce spurious foreground segments from other regions, one can confine the analysis within a region of interest (ROI), which can be determined manually or following a foreground accumulation approach [116]. Secondly, different scenarios may demand different background extraction strategies. Specifically, dynamic background subtraction [154] can cope with gradual illumination change but have difficulty in isolating people that are stagnant for a long period of time; static background subtraction [112,143] is able to segment static objects from the background but is susceptible to lighting change. Finally, poor estimation is expected if one employs only foreground area due to inter-object occlusion. Enriching the representation with other descriptors may solve this problem to certain extent.

2.2.2 Structural-Based Edge Features

While foreground features capture the global properties of the segment, edge features inside the segment carries complementary information about the local and internal patterns [19, 40, 87]. Intuitively, segments tend to present complex edges for inter-object occlusion occurs. Edges can be detected using an edge detector such as the Canny edge detector [17] with an example shown in Figure 2.4(c). Note that an edge image is often masked using the foreground segment to discard irrelevant edges. Some common edge-based features are listed as follows

- Total edge pixels – total number of edge pixels.
- Edge orientation – histogram of the edge orientations in the segment.
- Minkowski dimension – the Minkowski fractal dimension or box-counting dimension of the edges [119], which counts how many pre-defined structuring elements are required to fill the edges.

2.2.3 Local Texture Features

Textural information are also important visual cues in computer vision and a number of texture features were proposed such as gray-level co-occurrence matrix (GLCM) [68], local binary pattern (LBP) [131], HOG feature [38, 117], and gradient orientation co-occurrence matrix (GOCM) [117]. A comparative studies among the aforementioned texture and gradient features can be found in [117]. Here we provide a brief description on GLCM and LBP, which are widely used owing to its simple implementation and efficient computation.

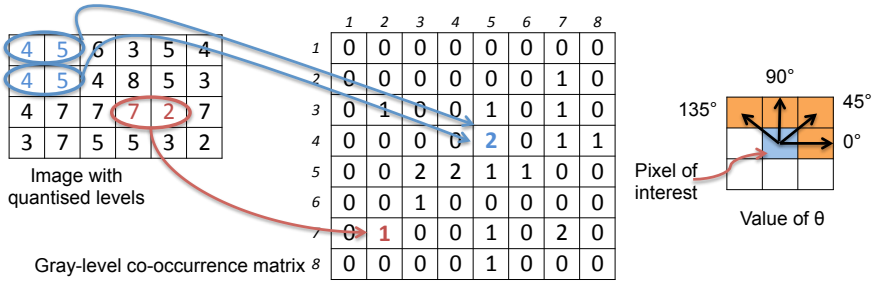


Figure 2.2: Gray-level co-occurrence matrix, with $\theta = 0°$ of a 4-by-6 image. Element (7,2) in the GLCM contains the value 1 because there is only one instance in the image where two, horizontally adjacent pixels have the values 7 and 2. Element (4,5) in the GLCM contains the value 2 because there are two instances in the image where two, horizontally adjacent pixels have the values 4 and 5. The value of θ specifies the angle between the pixel of interest and its neighbour.

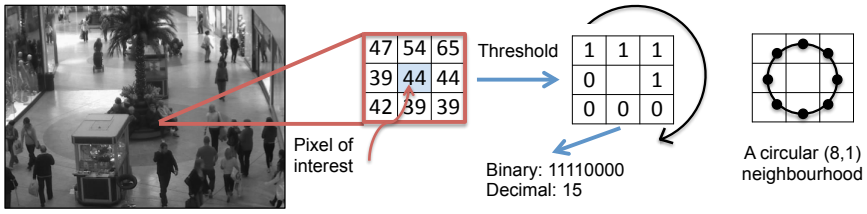


Figure 2.3: A basic local binary pattern operator [131] and a circular (8,1) neighbourhood.

Gray-level co-occurrence matrix (GLCM) – Gray-level co-occurrence matrix (GLCM) [68] is widely used in various computer vision problems such as crowd counting studies [19, 117, 120, 180]. To obtain GLCM, a typical process is to first quantise the image into 8 gray-levels and masked by the foreground segment. The joint probability or co-occurrence of neighbouring pixel values, $p(i, j \mid \theta)$ is then estimated for four orientations, $\theta \in \{0°, 45°, 90°, 135°\}$. After extracting the co-occurrence matrix, a set of features such as homogeneity, energy, and entropy can be derived for each θ

- Homogeneity – texture smoothness, $g_\theta = \sum_{i,j} \frac{p(i,j \mid \theta)}{1+|i-j|}$
- Energy – total sum-squared energy, $e_\theta = \sum_{i,j} p(i, j \mid \theta)^2$
- Entropy – texture randomness, $h_\theta = \sum_{i,j} p(i, j \mid \theta) \log p(i, j \mid \theta)$

Local Binary Pattern (LBP) – An alternative texture descriptor is the local binary pattern (LBP) [131]. Local binary pattern has been widely adopted in various applications such as face recognition [3] and expression analysis [147], due to its high discriminative power, invariance to monotonic gray-level changes, and its computational efficiency. An illustration of a basic

LBP operator is depicted in Figure 2.3. The LBP operation is governed by a definition of local neighbourhood, i.e. the number of sampling point and radius centering the pixel of interest. An example of a circular (8,1) neighbourhood is shown in Figure 2.3. Following the definition of neighbourhood, we sample 8 points at a distance of radius 1 from the pixel of interest and threshold them using the value of the centering pixel. The results are concatenated to form a binary code as the label of the pixel of interest. These steps are repeated over the whole image space and a histogram of labels is constructed as a texture descriptor. In this study, we employed an extension of the original LBP operator known as *uniform patterns* [131]. A uniform LBP pattern is binary code with at most two bitwise transitions, e.g. 11110000 (1 transition) and 11100111 (2 transitions) are uniform, whilst 11001001 (4 transitions) is not. In the construction of LBP histogram, we assign a separate bin for every uniform pattern and keep all nonuniform patterns in a single bin, so we have a 58+1-dimension texture descriptor.

2.2.4 Shape Features

Shape feature is also an important visual cue in computer vision, especially the applications for facial/human images analysis [50, 126] and pedestrian detection [38].

Scale-Invariant Feature Transform (SIFT) Features – Scale-invariant feature transform proposed by David Lowe [39] is an representation in computer vision to detect and describe local features in images. Specifically, extracting SIFT features has the following key stages: interest points localisation and local description. For interest points localisation, two main methods are adopted, i.e. pixel/grid-based and bag-of-words. At each interest points, a 128-d SIFT feature consisted of 4×4 cells and 8 orientation bins is obtained. A codebook is then generated by clustering the SIFT features into several categories. In this way, a SIFT feature can be represented by a cluster id rather than a 128-d descriptor. For different interest points localisation methods, the level of features can be different. On one hand, for pixel/grid-based SIFT features, location information of each pixel/grid is included implicitly, which can be viewed as a local-level feature representation. On the other hand, bag-of-words SIFT feature is a global-level feature missing the important location information. However, for extracting randomly-selected interest points SIFT features, the whole image space can be divided into several regions and then the histogram of each region can be concatenated together [28]. In addition to its robustness to the changes in illumination, noise, and minor changes in viewpoint, SIFT feature based on the appearance of

(a) Active Shape Model

(b) Shape-Free Texture Patches

(c) Active Appearance Model

Figure 2.4: An illustrative example of Active Appearance Model from FG-NET dataset [29].

the object at particular interest points is invariant to image scale and rotation in comparison with other features.

2.2.5 Hybrid Features

Active Appearance Model proposed by Cootes *et al* [35] is a statistical model to capture both shape and texture information via a set of training images and the coordinates of landmarks. AAM feature has been widely used in the applications with facial images such as facial age estimation [51, 64], face verification [134] and face recognition [191]. The model was exploited by combining a shape model and a texture model with using shape-free patches. Specifically, with the labelled landmarks of the facial images, a shape model can be learnt via applying Principal Component Analysis (PCA) to the coordinates of training samples. For testing a new image, a mean shape will be used and adjusted to fit local neighbouring. Although Active Shape Model (ASM) [36] works efficiently, the model cannot incorporate the gray-level textual information

and also may not converge to a good solution. In view of this, a texture model is developed by applying PCA to the gray-level patches matching the mean shape. For combining both shape and texture models together, another PCA is applied to concatenated shape and gray-level parameters. Benefiting from using all the information available, Active Appearance Model [35] is able to achieve more robust interpretation than Active Shape Model [36]. Illustrative Examples are given in Figure 2.4.

2.3 Geometric Correction

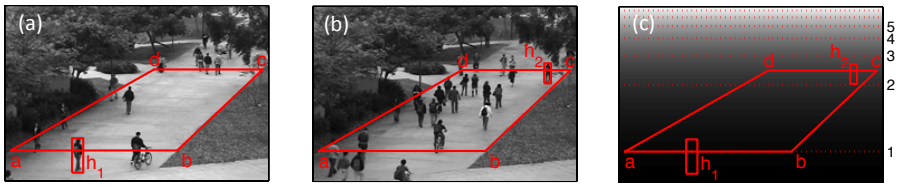
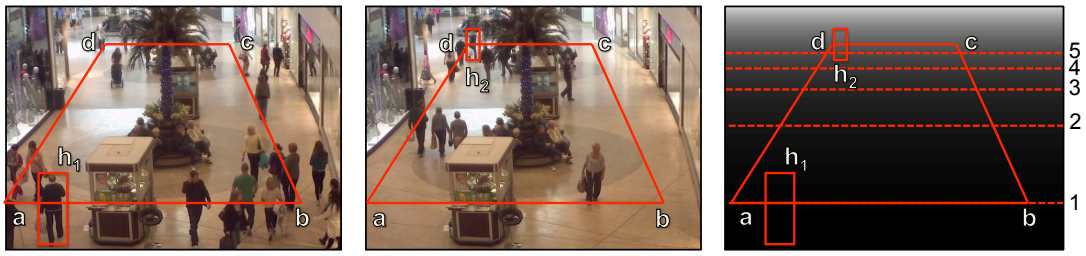


Figure 2.5: A perspective map in Figure (c) is generated through selecting a reference person at two extremes of a predefined quadrilateral as illustrated in Figures (a) and (b). Images from [19].

Before learning the regression models using the aforementioned feature representation, the problem of perspective distortion existing in crowd counting can lead to large variation of features, in which far objects appear smaller than those closer to the camera view. As a consequence, features (e.g. segment area) extracted from the same object at different depths of the scene would have huge difference in values. The influence is less critical if one divides the image space into different cells, each of which modelled by a regression function; erroneous results are expected if one only uses a single regression function for the whole image space.

To address this problem geometric correction or perspective normalisation is performed to bring perceived size of objects at different depths to the same scale. Ma et al [116] investigate the influence of perspective distortion to people counting and propose a principled way to integrate geometric correction in pixel counting, i.e. to scale each pixel by a weight, with larger weights given to further objects.

A simple and widely adopted perspective normalisation method [101, 108, 144] is described in [19]. The method first determines four points in a scene to form a quadrilateral that corresponds to a rectangle (see Figure 2.6) . The lengths of the two horizontal lines of the quadrilateral, $\overline{ab}$ and $\overline{cd}$, are measured as w_1 and w_2 respectively. When a reference pedestrian passes the two extremes, i.e. its bounding box's centre touches the $\overline{ab}$ and $\overline{cd}$, its heights are recorded



(a) (b) (c)

Figure 2.6: (a) and (b) show a reference person at two extremes of a predefined quadrilateral; (c) a perspective map to scale pixels by their relative size in the three-dimensional scene.

as h_1 and h_2. The weights at $\overline{ab}$ and $\overline{cd}$ are then assigned as 1 and $\frac{h_1 w_1}{h_2 w_2}$ respectively. To determine the remaining weights of the scene, linear interpolation is first performed on the width of the rectangle, and the height of the reference person. A weight at arbitrary image coordinate can then be calculated as $\frac{h_1 w_1}{h' w'}$, where h' and w' representing the interpolants. Here we make an assumption that the horizontal vanishing line to be parallel to the image horizontal scan lines. When applying the weights to features, it is assumed that the size of foreground segment changes quadratically, whilst the total edge pixels changes linearly with respect to the perspective. Consequently, each foreground segment pixel is weighted using the original weight and the edge features are weighted by square-roots of the weights.

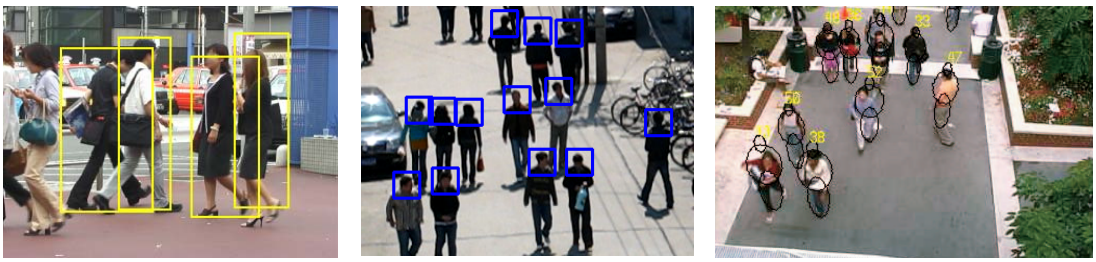
The aforementioned method [19] requires manual measurement which could be error-prone. There exist approaches to compute camera calibration parameters based on accumulative visual evidence in a scene. For example, a method is proposed in [89] to find the camera parameters by exploiting foot and head location measurements of people trajectories over time. Another more recent method [109] relaxes the requirement of accurate detection and tracking. This method takes noisy foreground segments as input to obtain the calibration data by leveraging the prior knowledge of the height distribution.

2.4 Crowd Density Estimation

Crowd counting in public places has a wide spectrum of applications especially in crowd control, public space design, and pedestrian behaviour profiling. Specifically, the problem of crowd counting is to estimate the exact number of persons within the whole video frames via different types of visual cues. The taxonomy of crowd counting algorithms can be generally grouped into three paradigms, namely counting by detection, clustering, and regression. In this section, we provide an overview on each of the paradigms, with a particular focus on the counting by

regression strategy that has shown to be effective on more crowded environments.

2.4.1 Counting by Detection



(a) (b) (c)

Figure 2.7: Pedestrian detection results obtained using (a) monolithic detection, (b) part-based detection, and (c) shape matching. Images from [100, 105, 195].

The following is a concise account of pedestrian detection with emphasizing on counting application. A more detailed treatment on this topic can be found in [42].

Monolithic detection: The most intuitive and direct approach to numerate the number of people in a scene is through detection. A typical pedestrian detection approach is based on monolithic detection [38, 100, 166], which trains a classifier using the full-body appearance of a set of pedestrian training images (see Figure 2.7(a)). Common features to represent the full-body appearance include Haar wavelets [169], gradient-based features such as histogram of oriented gradient (HOG) feature [38], edgelet [178], and shapelets [145]. The choice of classifier imposes significant impact on the speed and quality of detection, often requiring a trade-off between these two. Non-linear classifiers such as RBF Support Vector Machines (SVMs) offer good quality but suffer from low detection speed. Consequently, linear classifiers such as boosting [170], linear SVMs, or Random/Hough Forests [54] are more commonly used. A trained classifier is then applied in a sliding window fashion across the whole image space to detect pedestrian candidates. Less confident candidates are normally discarded using non-maximum suppression, which leads to final detections that suggest the total number of people in a given scene. Whole body monolithic detector can generates reasonable detections in sparse scenes. However, it suffers in crowded scenes where occlusion and scene clutter are inevitable [42].

Part-based detection: A plausible way to get around the partial occlusion problem to some extent is by adopting a part-based detection method [47, 107, 179]. For instance, one can construct boosted classifiers for specific body parts such as the head and shoulder to estimate the people

counts in a monitored area [105] (see Figure 2.7(b)). It is found that head region alone is not sufficient for reliable detection due to its shape and appearance variations. Including the shoulder region to form an omega-like shape pattern tends to give better performance in real-world scenarios [105]. The detection performance can be further improved by tracking validation, i.e. associating detections over time and rejecting spurious detections that exhibit coherent motion with the head candidates [133]. In comparison to monolithic detection, part-based detection relaxes the stringent assumption about the visibility of the whole body, it is thus more robust in crowded scenes.

Shape matching: Zhao *et al* [195] define a set of parameterised body shapes composed of ellipses, and employ a stochastic process to estimate the number and shape configuration that best explains a given foreground mask in a scene. Ge and Collins [56] extend the idea by allowing more flexible and realistic shape prototypes than just simple geometric shapes proposed in [195]. In particular, they learn a mixture model of Bernoulli shapes from a set of training images, which is then employed to search for maximum a posteriori shape configuration of foreground objects, revealing not only the count and location, but also the pose of each person in a scene.

Multi-sensor detection: If multiple cameras are available, one can further incorporate multi-view information to resolve visual ambiguities caused by inter-object occlusion. For example, Yang *et al* [184] extracted the foreground human silhouettes from a network of cameras to establish bounds on the number and possible locations of people. In the same vein, Ge and Collins [55] estimate the number of people and their spatial locations by leveraging multi-view geometric constraints. The aforementioned methods [55, 184] are restricted since a multi-camera setup with overlapping views is not always available in many cases. Apart from detection accuracy improvement, the speed of detection can benefit from the use of multi-sensors, e.g. the exploitation of geometric context extracted from stereo images [10].

Transfer learning: Applying a generic pedestrian detector to a new scene cannot guarantee satisfactory cross-dataset generalisation [42], whilst training a scene-specific detector for counting is often laborious. Recent studies have been exploring the transfer of generic pedestrian detectors to a new scene without human supervision. The key challenges include the variations of viewpoints, resolutions, illuminations, and backgrounds in the new environment. A solution to the problem is proposed in [173, 174] to exploit multiple cues such as scene structures, spatio-

(a) (b) (c)

Figure 2.8: (a) and (b) show the results of clustering coherent motions using methods proposed in [137] and [16] respectively. (c) shows the pairwise affinity of patches (strong affinity = magenta, weak affinity = blue) in terms of motion and colour constancy; the affinity is used to determine the assignment of patches to person hypotheses [165]. Images from [16, 137, 165].

temporal occurrences, and object sizes to select confident positive and negative examples from the target scene to adapt a generic detector iteratively.

2.4.2 Counting by Clustering

The counting by clustering approach relies on the assumption that individual motion field or visual features are relatively uniform, hence coherent feature trajectories can be grouped together to represent independently moving entities. Studies that follow this paradigm include [137], which uses a Kanade-Lucas-Tomasi (KLT) tracker to obtain a rich set of low-level tracked features, and clusters the trajectory to infer the number of people in the scene; and [16], which tracks local features and groups them into clusters using Bayesian clustering. Another closely related method is [165], which incorporates the idea of feature constancy into a counting by detection framework. The method first generates a set of person hypotheses of a crowd based on head detections. The hypotheses are then refined iteratively by assigning small patches of the crowd to the hypotheses based on the constancy of motion fields and intra-garment colour (see Figure 2.8(c)).

The aforementioned methods [16, 137] avoid supervised learning or explicit modelling of appearance features as in the counting by detection paradigm. Nevertheless, the paradigm assumes motion coherency, hence false estimation may arise when people remaining static in a scene, exhibiting sustained articulations, or two objects sharing common feature trajectories over time. Note that counting by clustering only works with continuous image frames, not static images whilst the counting by detection and regression do not have this restriction.

	Year	Features								Learning		Datasets
		segment	edge	texture	shape	intensity	gradients	motion	others	regression method	level	
Davies et al.[40]	1995	✓	✓	–	–	–	–	–	–	Linear regression	global	–
Marana et al.[121]	1997	–	–	✓	–	–	–	–	–	Self-organising map neural network	global	–
Cho et al.[32]	1997	✓	✓	–	–	–	–	–	–	Feedforward neural network	global	–
Kong et al. [87, 88]	2005 2006	✓	✓	–	–	–	–	–	–	Feedforward neural network	global	–
Dong et al.[43]	2007	–	–	–	✓	–	–	–	–	Shape matching + locally-weighted regression	segment	USC Campus Plaza
Chan et al.[19,20, 23]	2008 2009	✓	✓	✓	–	–	–	–	–	Gaussian processes	global	UCSD Pedestrian, PETS 2009
Chan et al.[22]	2009	✓	✓	✓	–	–	–	–	–	Bayesian Poisson regression	global	UCSD Pedestrian
Ryan et al.[144]	2009	✓	✓	–	–	–	–	–	–	Feedforward neural network	segment	UCSD Pedestrian
Cong et al.[33]	2009	✓	✓	–	–	–	–	–	–	Polynomial regression	segment	–
Lempitsky et al.[101]	2010	✓	–	–	–	✓	✓	–	–	Density function minimisation based on Maximum Excess over Subarrays distance	pixel	UCSD Pedestrian
Conte et al.[34]	2010	–	–	–	–	–	–	–	number of SURF points	Support vector regression	segment	PETS 2009
Benabbas et al.[9]	2010	✓	–	–	–	–	–	✓	–	Linear regression	segment	PETS 2009
Li et al.[104]	2011	✓	✓	–	–	–	–	–	–	Pedestrian detector + Linear regression	segment	CASIA Pedestrian [103]
Lin et al.[108]	2011	✓	✓	–	–	–	✓	–	–	Gaussian processes	segment	UCSD Pedestrian, PETS 2009
Chen et al.[30]	2012	✓	✓	✓	–	–	–	–	–	Ridge regression	segment	UCSD Pedestrian, Mall

Table 2.1: A table summarising existing counting by regression methods. Note that only publicly available datasets are listed in the datasets column [114].

2.4.3 Counting by Regression

Despite the substantial progress being made in object detection [42] and tracking [189] in recent years, performing either in isolation or both reliably in a crowded environment remains a non-trivial problem. Counting by regression deliberately avoids actual segregation of individual or tracking of features but estimate the crowd density based on holistic and collective description of crowd patterns. Since neither explicit segmentation nor tracking of individual are involved, counting by regression becomes a feasible method for crowded environments where detection and tracking are severely limited intrinsically.

One of the earliest attempts in exploring the use of regression method for crowd density es-

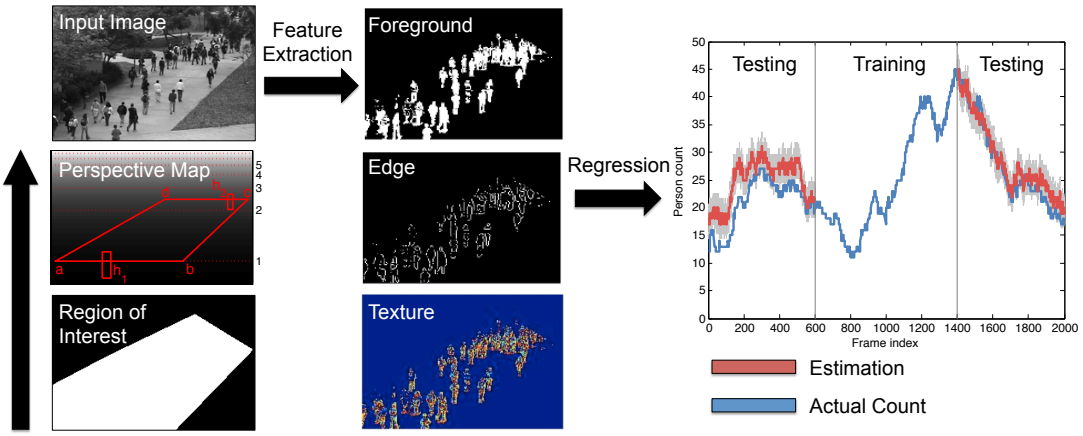


Figure 2.9: A typical pipeline of counting by regression: first defining the region of interest and finding the perspective normalisation map of a scene, then extracting holistic features and training a regressor using the perspective normalised features.

timation is by Davies et al [40]. They first extract low-level features such as foreground pixels and edge features from each video frame. Holistic properties such as foreground area and total edge count are then derived from the raw features. Consequently, a linear regression model is used to establish a direct mapping between the holistic patterns and the actual people counts. Specifically, a function is used to model how the input variable (i.e. the crowd density) changes when the target variables (i.e. holistic patterns) are varied. Given an unseen video frame, conditional expectation of the crowd density can then be predicted given the extracted features from that particular frame. Since the work of Davies et al [40], various methods have been proposed following the same idea with improved feature sets or more sophisticated regression models, but still sharing a similar processing pipeline as in [40]. A summary of some of the notable methods is given in Table 2.1.

2.5 Facial Age Estimation

Most existing techniques for age estimation from facial images fall into two categories: estimation-by-classification [24, 25, 59, 60, 185] (including ranking methods as specifical case of classification), estimation-by-regression [27, 46, 58, 65, 66, 111, 115, 125, 171, 194] (note that, hybrid [64] of the classification and regression in this thesis will be categorised into estimation-by-regression as regression framework has more effect on the performance and classification method is used to locally adjust), with regression models being the most widely used. As the detailed review for facial age estimation before 2009 is presented in [51], in this section, we will give the review on

the recent literatures after 2009.

2.5.1 Estimation by Classification

Under the assumption of different age patterns for different age groups, a learning-to-rank based framework was firstly applied to predicting human ages in [185], but priori object-identity knowledge of facial images is needed, which can thus be viewed as personalised human age estimation. Evidently, the setting of personalised age estimation is invalid in real-world applications, which is limited to the age estimation of multiple images from the same person. Chang et al. [24, 25] proposed an ordinary rank hyperplane method in the multiple independent binary classification form, which can mitigate the sparsity problem of training data with the price of time-consuming training multiple models proportional to age groups. Recently, Geng et al. [59] proposed a multilinear learning model with missing data, which can capture the latent structural information of tensor and also cope with the challenges of sparse and imbalanced data.

2.5.2 Estimation by Regression

The existing estimation by regression techniques for age estimation focus on improving the model performance by introducing more robust feature such as BIF feature [65] or proposing better regression models [58, 64, 66, 111, 115, 125, 171, 194]. Specifically, Guo et al. [64] proposed a locally adjusted regression method to search local regions for adjusting. Different regression techniques such as Support Vector Regression [115], Random Regression Forest [125], and Bayesian model [171] have been applied to facial age estimation in regression frameworks with Active Appearance Model feature. Zhang et al. [194] proposed a multi-task wrapped Gaussian Process Regression for personalised age estimation that jointly learns personalized characteristics and common changes shared between people. Long [111] proposed metric learning to find the intrinsic variation trend of age data. All the aforementioned estimation by regression methods are to tackle with feature variation challenge. In the light of the sparse and imbalanced data distribution in the existing benchmarks, some efforts have been devoted to mitigate the suffering by introducing multi-label learning [58] and tensor learning [66].

2.6 Human Pose Estimation

Human pose estimation including head pose estimation [66, 164] as its sub-topic is a basic but remaining challenging problem in computer vision and has a wide spectrum of applications such

as initial states of motion tacking [158], gait analysis [153], action recognition [187], 3D object or camera tracking [102] and detecting abnormal behaviour in the public place [73, 172]. In details, human pose estimation is to find the configuration (i.e. position and orientation) of human body parts in the images. In this section, we will give an review about the approaches for human pose estimation in still images, which is more challenging especially in unconstrained environment [45, 49]. Generally, the techniques for human pose estimation in single images can be categorised into estimation-by-detection and estimation-by-regression.

2.6.1 Estimation by Detection

In this subsection, the techniques with a pre-defined human geometric model [123, 124] for body part gesture estimation are presented as the following paragraphs. Model-based human pose estimation approaches specify a rough approximation of the skeleton and then use such a model in conjunction with image measurements to estimate the pose that best fits the model and the observed image features [95]. Those estimation-by-detection techniques are characterised by a kinematic model that relates constraints between body parts including kinematic constraints of articulated human as well as other constraints such as appearance constraints. Note that, body-parts localisation and pose estimation share similar characteristics, which can be learnt jointly to get benefits from each single task [76, 186].

Segment assembling – Segment assembling [140] was proposed by Ren et al. according to assembling the detected segments into the configuration of human body via pair-wise geometric constraints, i.e. edge. The concept for segment assembling is easy: instead of training a top-up body parts detector to find the candidate positions of body parts in the whole image space, edges in the image are selected and assembled according to the assumption that body parts in the image can be presented by a pair of parallel lines. It is worth pointing out that, segment assembling technique is similar to part-based estimation-by-detection methods, which can be viewed as the simple model with edge information only.

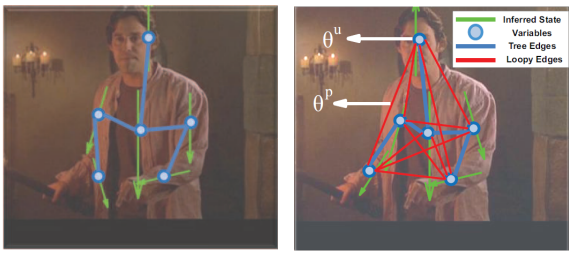
Shape matching – The configuration of human gesture can also be estimated via shape context matching [126]. Specifically, shape matching technique is based on the basic idea about matching testing shape with a number of exemplar human poses from different views. The limitation of such a techniques lies as the following. Firstly, the environment of images cannot be cluttered, which leads to a robust shape descriptor based on edges. Apparently, such a shape matching

Figure 2.10: Illustrative examples with part-based image parsing [138]. Images from [45, 49].

method cannot be applied in real-world applications, of which the environment is cluttered and out of control. Moreover, the number of exemplar posture in different views and the quality of labelled body joints have an effect on the performance, which is either impractical or labourious. More importantly, such a exemplar-based method has to suffer from the variations of appearance of human body in addition to human posture.

Skin colour detection – As presented in aforementioned paragraph of part-based estimation-by-detection, colour of the human body is important, e.g. clothing, skin. As the priori knowledge of the colour of clothing is unknown (although colour in different body parts has a strong correlation as [44]), skin colour is only available visual cues, which can be segmented directly from the images. In [74, 97, 98], a head/face detector was first applied to the image to locate the head position and then a skin-colour detector is learnt to find the regions with similar colour in the images, which will provide the strong evidence for human body configuration especially for the upper arms of scorer.

Part-based models – Part-based model for human pose estimation was firstly introduced in [47] and has been extended to improve the robustness of part detectors in a tree-based structure with combining both multiple visual appearance information (i.e. edge, colour) [44, 45, 49, 138, 157]. In brief, the part-based models are to learn the likelihood of each pixel in the foreground among different body parts constrained by relative correlation between pair-wise body parts. Considering the unknown prior knowledge of background and appearance of human body parts, the first work in this direction was proposed by Ramanan [138], which employed conditional random field in the generative manner with recursively updating human model by generic features (i.e. edges). Ferrari et al.proposed an algorithm to progressively reduce the search space with two pre-processing steps: human detection and foreground segmentation based on Grabcut [45, 49], as shown in Figure 2.10. Recent works are focusing on adding more priori knowledge or developing a more robust and informative feature representation. On one hand, in [44], more detailed prior



(a) Tree Model (b) Loopy Model

Figure 2.11: Comparison between tree models and loopy models. Images from [157].

knowledge was added to Ferrari's model with considering the stability of some body parts (e.g. torso, head) and statistically relation between different body parts. Furthermore, additional pose priors were combined with Pictorial Structures to achieve better performance [82]. On the other hand, sophisticated image features were exploited for further improving the performance of detecting body parts based on gradient [6,81,150] or color segmentation [80]. However, these tree-based graphical probabilistic graphical models could cause double-counting problem [138]: i.e. two independent localised legs might be recognised into the same body parts. To overcome the limitations of such a problem, the method incroporating loopy constraints by capturing pair-wise interactions of body parts was proposed with the price of approximate inference strategies [175]. Figure 2.11 illustrates the comparison between tree models and loopy models. In [186], both spatial relations between part locations and co-occurrence relations between part mixtures are combined together, which can decompose the a large set of global mixtures into local mixtures with capturing the global geometry by local appearance. Recently, a branch-and-bound algorithm was introduced into generative loopy graphical models to estimate human posture more efficiently [157].

2.6.2 Estimation by Regression

Different from estimation-by-detection techniques investigated in last section, human pose estimation can also be solved in a regression framework with learning a mapping between feature space and joint position output space. Generally, estimation-by-regression methods in a discriminative manner do not need the pre-defined human models and also benefit from the efficient inference algorithms in comparison with generative methods [44,45,49,138,157]. In this subsection, the estimation-by-regression methods for human posture will be categorised into supervised learning methods and semi-supervised learning methods.

Supervised learning models – A number of regression frameworks [2, 28, 61, 66, 76, 132, 156, 162, 167] for human pose estimation are developed in a supervised-learning fashion. In general, supervised estimation-by-regression methods are to learn the mapping between feature space and body configuration space with training images and labelled ground truth. In [2], a nonlinear regression framework based on Relevance Vector Regression is trained to capture histogram-of-shape-context features and human body configuration. However, regression-based frameworks for human pose estimation is made difficult due to the capturing of multiple highly-correlated joint output and ambiguous feature representation. In view of this, more efforts are dedicated to either more robust and informative low-level/intermediate feature representation [66, 132, 162] or more discriminative models [13, 28, 61, 66, 76, 149, 156, 167]. On one hand, for more robust feature representation, Okada et al. [132] employed histograms of orientated gradients [38] for feature selection within each manifolds. In [162], latent spaces are learnt for imagery feature space and pose space independently and then Gaussian Mixture Regression are employed for the mapping between these two latent spaces. Tensor feature representation for human pose estimation is proposed in [66], which is learning the mapping between tensor features to each joint output independently. On the other hand, the structural information between human body joints are critical and informative. In the light of this, a number of literatures focused on exploiting a unique framework for learning the correlation between output elements and input feature. In [13, 28, 76], structured output learning based algorithms (e.g. Structural Support Vector Machines) were proposed for capturing the mapping not only between input imagery feature vector and output but also each element of output vector. In [61], a regression model was proposed to learn the configuration of human body from the depth images directly. Shotton et al. [149] then introduced an intermediate representation from depth images in a random forest regression framework, which can achieve superior efficiency for human pose estimation in real-time. Recently, conditional regression forests [156] was applied to the problem of human pose estimation with incorporating conditional dependency on output variables (e.g. the height of the person to be estimated) on the basis of Shotton's work.

Semi-supervised learning models – In supervised learning regression frameworks for human pose estimation [2, 28, 61, 66, 76, 132, 156, 162, 167], there exists the assumption that sufficient training samples are available. Evidently, a large nubmer of images can be found easily from the Internet, but labelling human poses in the images is labourious and time-consuming. Semi-

supervised learning models [14, 83, 128, 136] relax the assumption and achieve the values in the practical viewpoint. In particular, semi-supervised frameworks [14, 128] with manifold learning are proposed to overcome the lackness of training data in joint/structual learning manner. For further mining the manifold structure of data, graph-based models [83, 136] were proposed by working on the localities.

2.7 Regression Models

After feature extraction and constructing model input and output, a regression model is selected and trained to predict the continuous labels given the normalised features. A regression model may have a broad class of functional forms. In this section we will review a few popular regression models in computer vision.

2.7.1 Linear regression

Given a training data comprising N observations $\{\mathbf{x}_n\}$, where $n = 1, 2, \ldots, N$ together with corresponding continuous target values $\{y_n\}$, the goal of regression is to predict the value of y given a new value of $\mathbf{x}$ [11]. The simplest approach is to form of linear regression function $f(\mathbf{x}, \mathbf{w})$ that involves a linear combination of the input variables, i.e.

$$f(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D, \tag{2.1}$$

where D is the dimension of features, $\mathbf{x} = (x_1, \ldots, x_D)^\mathsf{T}$, and $\mathbf{w} = (w_0, \ldots, w_D)^\mathsf{T}$ are the parameters of the model. This model is often known as *linear regression* (LR), which is a linear function of the parameters $\mathbf{w}$. In addition it is also linear with respect to the input variables $\mathbf{x}$.

To relax the linearity assumption, one can take a linear combination of a fixed set of nonlinear functions of the input variables, also known as basis functions $\phi(\mathbf{x})$, to obtain a more expressive class of function. It has the form of

$$f(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x}), \tag{2.2}$$

where M is the total number of parameters in this model, $\mathbf{w} = (w_0, \ldots, w_{M-1})^\mathsf{T}$, and $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^\mathsf{T}$. The functional form in (2.2) is still known as linear model since it is linear in $\mathbf{w}$, despite the function $f(\mathbf{x}, \mathbf{w})$ is nonlinear with respect to input vector $\mathbf{x}$. A polynomial re-

gression function is a specific example of this model, with the basis functions taking a form of powers of $\mathbf{x}$, that is $\phi_j(\mathbf{x}) = \mathbf{x}^j$. Gaussian basis function and sigmoidal basis function are other possible choices of basis functions.

Parameters in the aforementioned linear model are typically obtained by minimising the sum of squared errors

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2. \tag{2.3}$$

One of the key limitation of linear model is that the model can get unnecessarily complex given high-dimensional observed data $\mathbf{x}$. In addition, some of high-dimensional features may be highly co-linear, unstable estimate of parameters may occur [11], leading to very large magnitude in the parameters and therefore a clear danger of severe over-fitting.

2.7.2 Partial Least Squares Regression

A way of addressing the multicollinearity problem is by *partial least squares regression* (PLSR) [57], which projects both input $\mathbf{X} = \{\mathbf{x}_n\}$ and target variables $\mathbf{Y} = \{y_n\}$ to a latent space, with a constraint such that the lower-dimensional latent variables explain as much as possible the covariance between $\mathbf{X}$ and $\mathbf{Y}$. Formally, the PLSR decomposes the input and target variables as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^\mathsf{T} + \varepsilon_x \tag{2.4}$$
$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^\mathsf{T} + \varepsilon_y, \tag{2.5}$$

where $\mathbf{T}$ and $\mathbf{U}$ are known as score matrices, with the column of $\mathbf{T}$ being the latent variables; $\mathbf{P}$ and $\mathbf{Q}$ are known as loading matrices [1]; and ε are the error terms. The decompositions are made so to maximise the covariance of $\mathbf{T}$ and $\mathbf{U}$. There are two typical ways in estimating the score matrices and loading matrices, namely NIPALS and SIMPLS algorithms [1, 188].

2.7.3 Kernel Ridge Regression

Another method of mitigating the multicollinearity problem is through adding a regularisation term to the error function in Equation (2.3). A simple regularisation term is given by the sum-of-squares of the parameter vector elements, $\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w}$. The error function becomes

$$E_R(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{y_n - \mathbf{w}^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}_n)\right\}^2 + \frac{\lambda}{2}\mathbf{w}^\mathsf{T}\mathbf{w}, \tag{2.6}$$

with λ to control the trade-off between the penalty and the fit. A common way of determining λ is via cross-validation. Using this particular choice of regularisation term with $\phi(\mathbf{x}_n) = \mathbf{x}_n$, we will have error function of *ridge regression* (RR) [72].

A non-linear version of the ridge regression, known as *kernel ridge regression* (KRR) [146], can be achieved via kernel trick [148], whereby a linear ridge regression model is constructed in higher dimensional feature space induced by a kernel function defining the inner product

$$k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^\mathsf{T} \phi(\mathbf{x}'). \tag{2.7}$$

For the kernel function, one has typical choices of linear, polynomial, and radial basis function (RBF) kernels. The regression function of KRR is given by

$$f(\mathbf{x}, \boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n k(\mathbf{x}, \mathbf{x}_n), \tag{2.8}$$

where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}^\mathsf{T}$ are Lagrange multipliers. This solution is not sparse in the variables α, that is $\alpha_n \neq 0$, $\forall n \in \{1, \ldots N\}$.

2.7.4 Support Vector Regression

Support vector regression (SVR) [92, 151] has been widely used for different computer vision problems in [63,180]. In contrast to KRR, the SVR achieves sparseness in α (see Equation (2.8)) by using the concept of support vectors to determine the solution, which can result in faster testing speed than KRR that sums over the entire training-set [176]. Specifically, the regression function of SVR can be written as

$$f(\mathbf{x}, \boldsymbol{\alpha}) = \sum_{\text{SVs}} (\alpha_n - \alpha_n^*) k(\mathbf{x}, \mathbf{x}_n) + b, \tag{2.9}$$

where α_n and α_n^* represents the Lagrange multipliers, $k(\mathbf{x}, \mathbf{x}_n)$ denotes the kernel, and $b \in \mathbb{R}$. A popular error function for SVR training is the ε-insensitive error function [168], which assigns zero error if the absolute difference between the prediction $f(\mathbf{x}, \boldsymbol{\alpha})$ and the target y is less than $\varepsilon > 0$. *Least-squares support vector regression* (LSSVR) [160] is a least squares version of SVR. In LSSVR one finds the solution by solving a set of linear equations instead of a convex quadratic error function as in conventional SVR.

2.7.5 Gaussian Processes Regression

One of the most popular nonlinear methods is *Gaussian processes regression* (GPR) [139]. It has a number of pivotal properties – it allows a possibly infinite number of basis functions driven by the data complexity, and it models uncertainty in regression problems elegantly[1]. Formally, we write the regression function as

$$f(\mathbf{x}) \sim \mathrm{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \tag{2.10}$$

where Gaussian processes, $\mathrm{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is specified by its mean function $m(\mathbf{x})$ and covariance function or kernel $k(\mathbf{x}, \mathbf{x}')$

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})], & (2.11)\\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))], & (2.12) \end{aligned}$$

where $\mathbb{E}$ denotes the expectation value.

Apart from the conventional GPR, various extensions of it have been proposed for crowd counting. For instance, Chan et al. [18] propose a generalised Gaussian process model, which allows different parameterisation of the likelihood function, including a Poisson distribution for predicting discrete counting numbers [22]. Lin et al. [108] employ two GPRs in their framework, one for learning the observation-to-count mapping, and another one for reasoning the mismatch between predicted count and actual count due to occlusion. In age estimation, multi-task wrapped GPR [194] is proposed for personalised age estimation with an additional step of wrapping the estimated values.

The key weakness of GPR is its poor tractability to large training sets. Various approximation paradigms have been developed to improve its scalability [139].

It is worth pointing out that one of the attractive properties of kernel methods such as KRR, SVR, and GPR is the flexibility of encoding different assumptions about the function we wish to learn. For instance, by combining different covariance functions $k(\mathbf{x}, \mathbf{x}')$, such as linear, Matérn, rational quadratic, and neural network, one has the flexibility to encode different assumptions on the continuity and smoothness of the GP function $f(\mathbf{x})$. This property is exploited in [19], in which linear and a squared-exponential (RBF) covariance functions are combined to capture both the linear trend and local non-linearities in the crowd feature space.

[1] One can also estimate the predictive interval in other kernel methods such as KRR [41].

2.7.6 Random Forest Regression

Scalable nonlinear regression modelling can be achieved using *random forest regression* (RFR). A random forest comprises of a collection of randomly trained regression trees, which can achieve better generalisation than a single over-trained tree [37]. Each tree in a forest splits a complex nonlinear regression problem into a set of subproblems, which can be more easily handled by weak learners such as a linear model[2]. To train a forest, one optimises an energy over a given training set and associated values of the target variable. Specifically, parameters $\boldsymbol{\theta}_j$ of the weak learner at each split node j are optimised via

$$\boldsymbol{\theta}_j^* = \underset{\boldsymbol{\theta}_j \in \mathcal{T}_j}{\operatorname{argmax}} I_j, \tag{2.13}$$

where $\mathcal{T}_j \subset \mathcal{T}$ is a subset of parameters made available to the j-th node, and I is an objective function that often takes the form of information gain. Given a new observation $\mathbf{x}$, the predictive function is computed by averaging individual posterior distributions of all the trees, i.e.

$$f(\mathbf{x}) = \frac{1}{T} \sum p_t(y|\mathbf{x}), \tag{2.14}$$

where T is the total number of trees in the forest, $p_t(y|\mathbf{x})$ is the posterior of t-th tree.

The hallmark of random forest is its good performance comparable to state-of-the-art kernel methods (e.g. GPR), but with the advantage of being scalable to large dataset and less sensitive to parameters. In addition, it has the ability of generating variable importance and information about outliers automatically. It is also reported in [37] that regression forest can yield a more realistic uncertainty in the ambiguous feature region, in comparison to GPR that tends to return largely over-confident prediction.

2.8 Summary

Regression frameworks are widely employed for solving computer vision problems such as crowd counting, facial age estimation, and human pose estimation. According to the characteristics of different vision problems, we will face to address different problems with different strategies to improve the performance. However, the common nature of regression frameworks are unique, which can be readily transformed to other applications in regression frameworks, i.e.

[2]There are other weak learners that define the split functions, such as general oriented hyperplane or quadratic function. A more complex splitting function would lead to higher computational complexity.

exploiting more discriminative and robust feature representation to mining the spatial correlation between localised regions (applied to crowd counting problem in Chapter 3); introducing a novel attribute space to mitigate sparse and imbalanced data problem by capturing latent cumulative dependency in label space (applied to facial age estimation and crowd density estimation in Chapter 4); capturing the underlying structural information between each element of output vector (applied to human pose estimation in Chapter 5).

Chapter 3

Feature Mining for Localised Crowd Counting

(a) the UCSD Dataset [19] (b) the Mall Dataset [30]

Figure 3.1: (a) the UCSD benchmark dataset and (b) the Mall dataset.

Crowd counting in public places has a wide spectrum of applications especially in crowd control, public space design, and pedestrian behaviour profiling. As illustrated in Figure 3.1, examples of two public benchmark datasets for outdoor and indoor scenes are given. In some applications, e.g. crowd counting on a train platform, estimating a global count for the whole scene is sufficient. For more complex scenarios, it is necessary to estimate the counts at different spatial locations as well. For instance, for crowd counting in a shopping mall as 3.1(b) shows, one needs to know not only how many people in total are in the scene, but also where they are distributed, i.e. which shop is more popular.

Existing people counting techniques fall into three categories: counting by detection, counting by clustering, and counting by regression. As reviewed in previous chapter, the counting by detection [56, 105, 195] and by clustering [16, 137] approaches either rely on explicit object segmentation or feature point tracking. They are not suitable for crowded scenes with cluttered

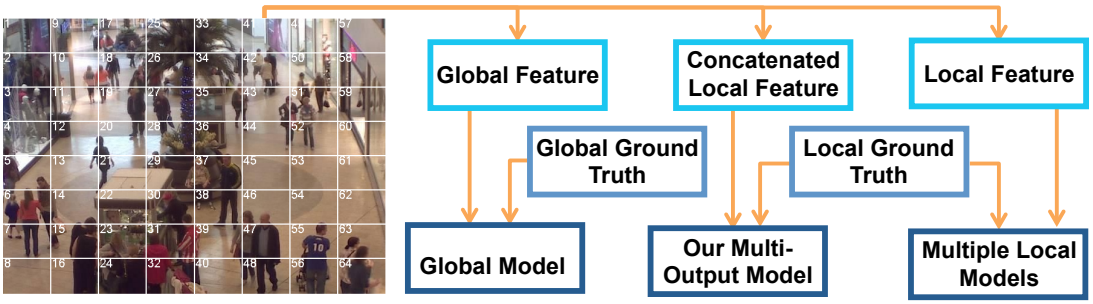


Figure 3.2: A flow chart illustrating the processing pipeline of global and local counting by regression methods, and our multi-output model.

background and frequent inter-object occlusion. In contrast, a counting by regression model aims to learn a direct mapping between low-level features and people count without segregation or tracking of individuals. This approach is more suitable for crowded environments and is computationally more efficient.

Existing counting by regression methods can be categorised into either global approaches or local approaches (see Figure 3.2). Global approaches [19, 23, 32, 87, 121, 144] learn a single regression function between image features extracted globally from the entire image space and the total people count in that image. Since spatial information is lost when computing global features, such a model assumes implicitly that a feature should be weighted the same regardless where in the scene it is extracted. However, this assumption is largely invalid in real-world scenarios. In particular, crowd structures[1] can vary spatially due to density, scene layout, and self-organisation of crowd induced by elementary individual interactions, boundary conditions, and regulations [71]. Thus, different features can be more reliable and relevant for crowd counting at different spatial locations. Furthermore, a global regression model is unable to provide information about spatially local crowd count information, which is desired.

To overcome these limitations of a global approach, local models [117, 180] aim to relax the global assumption to certain extent by dividing the image space into cell regions, each of which modelled by a separate regression function. The regions can be cells having regular size, or different resolutions determined by the scene perspective to compensate camera geometric distortions [117]. Local counts can be estimated in each region and a global count can then be obtained by summing up the cell-level counts. In an extreme case, Lempitsky et al. [101] go one step further to model the crowd density at each pixel, casting the problem as that of estimating

[1] Systematic granular motion of crowd resembling the flow of gas, fluid, and granular media.

an image density whose integral over any image region gives the count of objects within that region. In general, unlike global approaches [23, 32, 87, 121], local models aim to weigh features differently by local crowd structures in order to facilitate localised crowd counting. However, existing local methods suffer a scalability issue due to the need to learn multiple regression models, the number of which can become very large. In addition, an inherent drawback of existing local models is that no information is shared across spatially localised regions in order to provide a more context-aware feature selection for more accurate crowd counting. In many real-world cases, low-level imagery features can be highly ambiguous due to cluttered background and severe inter-object occlusions. Therefore, harnessing common properties and features among different local spatial regions should benefit the estimation of crowd density.

3.1 The Concept

We consider that *localised feature importance mining* and *information sharing among regions* are two key factors for accurate and robust crowd counting, which are missing in all existing techniques. To this end, we propose a single multi-output model for joint localised crowd counting based on ridge regression [146], which takes inter-dependent local features from local spatial regions as input and people count from individual regions as a multi-dimensional structured output (see Figure 3.2). Unlike global regression methods, our model relaxes the one-to-one mapping assumption by learning spatially localised regression functions jointly in a single model for all the individual cell regions in a scene, as such our model can capture feature importance locally. Unlike existing approaches to building multiple local regression models, our single model is learned by joint optimisation to enforce dependencies among cell regions. Therefore information from all local spatial regions can be shared to achieve more reliable count prediction. We demonstrate the effectiveness of our model on both an existing crowd analysis benchmark dataset and a new more challenging shopping mall dataset. In summary, the main contributions and novelties of this study are two-fold:

- This is the first study that achieves robust crowd counting by mining local feature importance and sharing visual information among spatially localised regions in a scene.
- This is achieved by considering a single multi-output ridge regression model for localised crowd counting which has advantages over both existing global approaches in providing local estimates and existing local approaches being more scalable.

3.2 Methodology

Figure 3.3 gives an overview of our framework:

- (Step-1) We first infer a perspective normalisation map using the method described in [19].
- (Step-2) Given a set of training images, we extract low-level imagery features, including local foreground, edges and texture features, from each cell region.
- (Step-3) Local features from each cell are used to construct a local intermediate feature vector before all local intermediate feature vectors are concatenated into a single ordered (location-aware) feature vector.
- (Step-4) A multi-output regression model based on multivariant ridge regression is trained using the single concatenated feature vector and the vector, each element being the actual count in each region, as a training pair.

Given a new test frame, features are extracted and mapped to the learned regression model for generating a structured output that estimates the crowd count in each local region simultaneously.

Note that the training/testing procedure adopted in our framework is similar to that in a global counting framework (see Figure 3.2), but with a different and new learning strategy to enable spatially localised features weighting and inter-region feature sharing. This variation is important to our approach. As in [101], our method requires dot annotations on each pedestrian so we can generate a training count for each cell region. This may appear a laborious task but dotting/pointing is the natural way of how human numerate objects, in practice dotted annotation is no harder than a raw count as in the global counting methods [23].

3.2.1 Feature Representation

Given a training video frame i, where $i = 1,2\ldots N$ and N denotes the total number of training frames, we first partition the frame into K cell regions (see Step-3 in Figure 3.3). We then extract low-level imagery features $\mathbf{z}_i^j$ from each cell region j and combine them into an intermediate feature vector $\mathbf{x}_i \in \mathbb{R}^d$. We also concatenate the localised labelled ground truth u_i^j from each cell region into a multi-dimensional output vector, $\mathbf{y}_i \in \mathbb{R}^m, i = 1,2\ldots N$

$$\mathbf{x}_i = [\mathbf{z}_i^1, \mathbf{z}_i^2, \ldots, \mathbf{z}_i^{K-1}, \mathbf{z}_i^K], \quad \mathbf{y}_i = [u_i^1, u_i^2, \ldots, u_i^{K-1}, u_i^K].$$

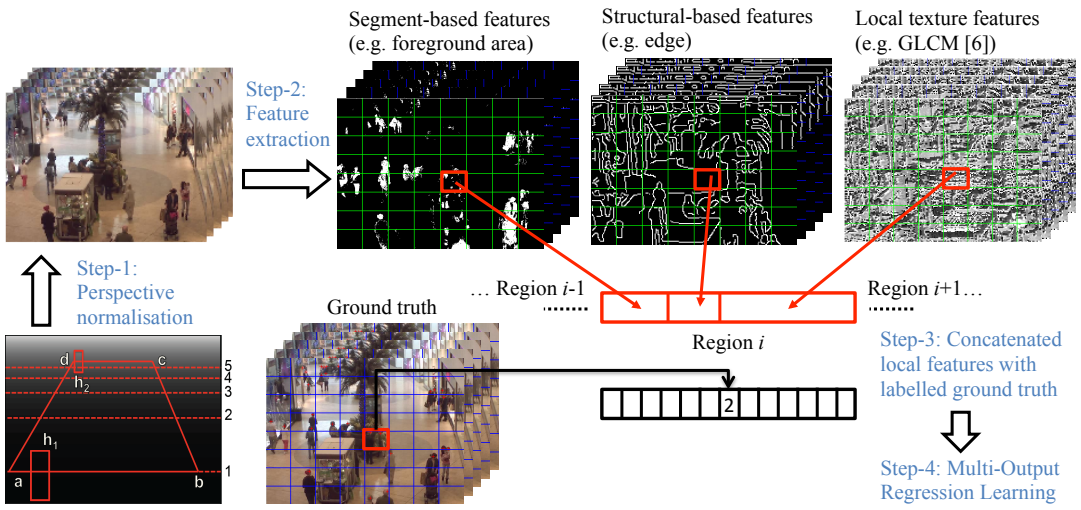


Figure 3.3: A multi-output regression framework for localised crowd counting by feature mining.

We train the proposed model using $\{(\mathbf{x},\mathbf{y})\}_i, i = 1,2\ldots N$. In this study, we adopt three types of features as in [19], which also be reviewed with more details in Section 2.2:

- Segment-based features (refer to Section 2.2.1): foreground area, total number of pixels of the foreground perimeter, perimeter-area ratio, histogram of perimeter edge orientation, and blob count ;
- Structural-based features (refer to Section 2.2.2): total number of edge pixels, histogram of the edge orientation, and Minkowski dimension;
- Local texture features (refer to Section 2.2.3): Gray Level Co-occurrence Matrix (GLCM) [68].

Note that all images are transformed to grayscale prior to feature extraction. In addition, features are perspective normalised using the method described in [19] and scaled into $[0,1]$.

Before ending this subsection, we are going to investigate the individual features used here and more details can be found in our work [114] as well as [193].

Robustness of individual features: It is observed that different features can be more important given different crowdedness levels. In general, the performance in [114] suggests that the segment-based features were superior to other features. This is not surprising since the foreground segment carries useful information about the area occupied by objects of interest and it thus intrinsically correlate to the number of pedestrians in a scene. However in a more crowded environment with frequent inter-object occlusion, segment-based features would suffer, whilst

edge and texture that inherently encoded the inter-object boundary and internal patterns would carry more discriminative visual cues for density mapping.

3.2.2 Multi-Output Regression Model

For learning a multi-output regression model, we exploit the ridge regression function [5, 67]. In its conventional form, a ridge regression function learns a single output mapping. In our case, we adapt it to cope with a multi-output regression learning problem for simultaneous localised crowd counting in different spatial cell regions. The rational for exploiting ridge regression is that the model offers superior robustness in coping with multicollinearity problem[2], due to its regularised least-square error minimisation, as opposed to ordinary least-square in classic regression methods such as linear regression. Ridge regression has been exploited elsewhere for face recognition [5]. This is the first attempt to exploit it for crowd analysis.

Formally, given ($\mathbf{x}_i$, $\mathbf{y}_i$) as the observation and target vectors, multivariate ridge regression can be presented as follows

$$\min \quad \left(\frac{1}{2} \|\mathbf{W}\|_F^2 + C \sum_{i=1}^{N} \|\mathbf{y}_i^T - \mathbf{x}_i^T \mathbf{W} - \mathbf{b}\|_F^2 \right), \tag{3.1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times K}$ and $\mathbf{b} \in \mathbb{R}^{1 \times K}$ denote a weight matrix and a bias vector respectively. $\|\cdot\|_F$ denotes the Frobenius-norm, and C is a parameter that controls the trade-off between the penalty and the fit.

The weight matrix $\mathbf{W}$ plays an important role in capturing the local feature importance and facilitating the sharing of features. In particular, for each localised cell, we formulate our model to jointly weigh the features extracted from both the corresponding localised cell and other cell regions in the image. According to the above Equation (3.1), for jth cell region in the images, jth column of matrix $\mathbf{W}$ is employed to weigh the concatenated feature vector $\mathbf{x}_i$ for the count estimation in corresponding localised cell region, i.e. jth entry of $\mathbf{y}_i$. Considering the residual error of all cell regions being penalized jointly with Frobenius-norm and feature vector $\mathbf{x}_i$ consisting of feature from both jth cell region and other cell regions in the image, such a regression model can benefit from local feature importance mining, and more importantly, feature information sharing within the whole image space for the localised crowd density estimation of a specific region.

[2]Some low-level features may be highly co-linear, unstable estimate of parameters may occurs [11], leading to very large magnitude in the parameters and therefore a clear danger of severe over-fitting.

Here we provide more details on the error minimisation. Specifically, the above Equation (3.1) is transformed as follows

$$\min \quad M(\theta) = \mathrm{tr}\left(\frac{1}{2}\theta^T \mathbf{Q}\theta + \mathbf{P}^T\theta\right), \tag{3.2}$$

where the positive semi-definite matrix Q and matrix P are given as

$$\mathbf{Q} = \begin{bmatrix} 2C\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T + I & 2C\sum_{i=1}^{N}\mathbf{x}_i \\ 2C\sum_{i=1}^{N}\mathbf{x}_i^T & 2CN \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \quad \mathbf{P} = \begin{bmatrix} -2C\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i^T \\ -2C\sum_{i=1}^{N}\mathbf{y}_i^T \end{bmatrix} \in \mathbb{R}^{(d+1)\times K},$$

where $\theta = [\mathbf{W};\mathbf{b}] \in \mathbb{R}^{(d+1)\times K}$ represents the matrix to be optimized, I denotes the identity matrix, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Different from the standard ridge regression with single output, the coefficient P and parameters θ to be optimized in Equation (3.2) are matrices instead of vectors, which leads to the usage of the trace $\mathrm{tr}(\cdot)$ for minimisation. Similar to ridge regression, Equation (3.2) is solved using the Quadratic Programming, which has a global optimal solution, if and only if

$$\frac{\partial M(\theta)}{\partial \theta} = \mathbf{Q}\theta + \mathbf{P} = \mathbf{0},$$

and thus, the weights and bias of ridge regression are computed by

$$\theta = -(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{P}.$$

An alternative to the multi-output ridge regression model is the structural Support Vector Machine [79], which has been applied to pose estimation [76] and object detection [12]. Multivariate ridge regression is adopted owing to its simplicity in implementation.

3.3 Experiments

To verify the effectiveness of our proposed framework for localised crowd counting and also show the insights of feature selected by our model, we will conduct the experiments with the following settings and analysis.

3.3.1 Datasets and Settings

Datasets – The effectiveness of the proposed method was evaluated on two datasets: the UCSD benchmark dataset [19, 23] and the Mall dataset [30]. The details of the two datasets are given

Data	N_f	R	FPS	D	Tp
UCSD [19]	2000	238 × 158	10	11–46	49885
Mall [30]	2000	320 × 240	<2	13–53	62325

Table 3.1: Dataset properties: N_f = number of frames, R = Resolution, FPS = frame per second, D = Density (minimum and maximum number of people in the ROI), and Tp = total number of pedestrian instances.

in Table 4.1, and the example frames are shown in Figure 3.1. In contrast to the UCSD dataset, of which the video was recorded from a campus scene using hand-held camera (Figure 3.1(a)), the Mall dataset (Figure 3.1(b)) was captured using a publicly accessible surveillance camera in a shopping mall with more challenging lighting conditions and glass surface reflections. The Mall dataset also covers more diverse crowd densities from sparse to crowded, as well as different activity patterns (static and moving crowds) under larger range of illumination conditions at different time of the day. In addition, in comparison to the UCSD dataset, the Mall dataset experiences more severe perspective distortion, which causes larger changes in size and appearance of objects at different depths of the scene, and has more frequent occlusion problem caused by the scene objects, e.g. stall, indoor plants along the walking path.

Settings – For the UCSD dataset, we followed the same training and testing partition as in [19], i.e. we employed Frames 601-1400 for training and the rest for testing. For the Mall dataset, we used the first 800 frames for training and kept the remaining 1200 frames for testing. Based on the resolution of the different datasets, we defined 6×4-cells for the UCSD dataset and 8×8-cells for the Mall dataset.

Evaluation Metrics – We employed three evaluation metrics, namely *mean absolute error* (mae), ε_{abs}; *mean squared error* (mse), ε_{sqr}; and *mean deviation error* (mde), ε_{dev}.

$$\varepsilon_{\text{abs}} = \frac{1}{N}\sum_{i=1}^{N}|v_i - \hat{v}_i|, \quad \varepsilon_{\text{sqr}} = \frac{1}{N}\sum_{i=1}^{N}(v_i - \hat{v}_i)^2, \quad \text{and} \quad \varepsilon_{\text{dev}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|v_i - \hat{v}_i|}{v_i},$$

where N is the total number of test frames, v_i is the actual count in each cell region or the whole image, and $\hat{v}_i$ is the estimated count of ith frame.

3.3.2 Comparative Evaluation

We compared the following models

- Single global model with global feature (1) ridge regression (RR) [146], and (2) Gaussian processes regression (GPR) with linear + RBF kernel as in [19]. These models employ

Method	Features Level		Learning Level		UCSD [19]			Mall		
	Global	Local	Global	Local	mae	mse	mde	mae	mse	mde
RR [146]	√	–	√	–	2.25	**7.82**	0.1101	3.59	19.0	0.1109
GPR [19]	√	–	√	–	**2.24**	7.97	0.1126	3.72	20.1	0.1159
MLR [180]	–	√	–	√	2.60	10.1	0.1249	3.90	23.9	0.1196
MORR	–	√	√	–	2.29	8.08	**0.1088**	**3.15**	**15.7**	**0.0986**

Table 3.2: Performance comparison between different methods and our multi-output ridge regression (MORR) model on global crowd counting. Note that, the results in this table are based on our implementation.

global features as their input and the crowd density of the whole image as their output.

- Multiple localised regressors (MLR) [117]. The input and output for the model is the feature within each cell and the people count in the corresponding cell respectively. The ridge regression model is used to eliminate the effect of using different regression models.
- The proposed multi-output ridge regression (MORR) model described in Sec. 3.2.2.

For all models free parameters were tuned using 4-fold cross-validation.

3.3.3 Comparison With Single Global Regression Models

The results of different models on the two datasets are shown in Table 3.2. It can be observed the two global regression models, RR and GPR, yielded very similar results on UCSD compared to our MORR model, but much higher error rates on the more challenging Mall dataset, i.e., 16.03%, 24.52%, and 15.01% higher than our model in mae, mse, and mde on average. It is worth pointing out that in contrast to the other two metrics the mde is more indicative as it takes the level of crowdedness of ith frame into account.

Different performances on two datasets were due to the different characteristics of the two scenes. In the shopping mall scene, different local regions can have drastically different lighting conditions (see Figure 3.1(b)). The different fixed structures in the scene (e.g. stalls and plants in the middle) also introduced different characteristics of occlusion. In comparison, in the UCSD campus scene, there was no occlusion caused by static objects and the lighting condition across the scene were fairly even and stable during the entire recording period.

The result thus suggests that mining features at different spatial location is more critical for a complex scene where lighting conditions are not uniform and can change quickly, and occlusion can occur both inter people and between people and static obstacles. It is also worth pointing out that despite its simpler formulation, the single global ridge regression model achieves comparable

or better performance compared to the more complex Gaussian Processes Regression model.

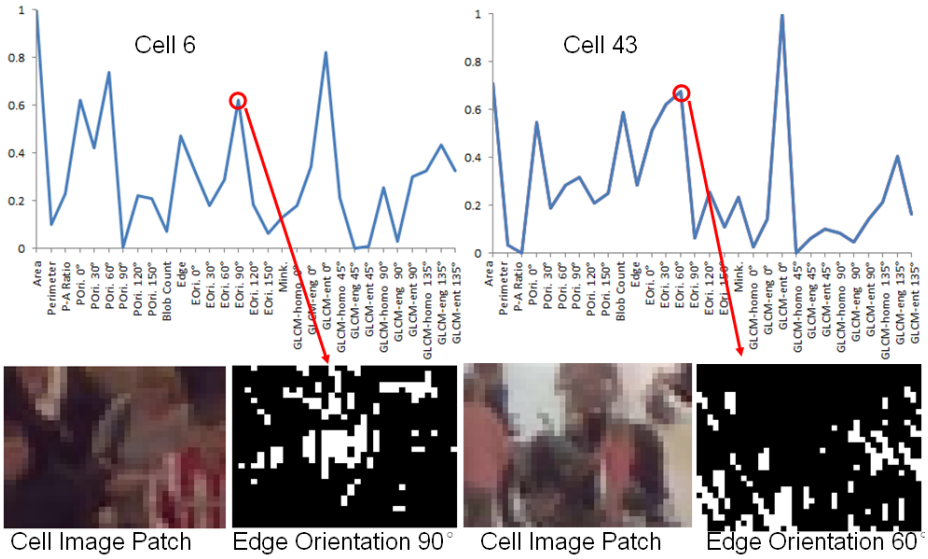


Figure 3.4: Local feature mining from one Close-to-Camera Cell 6 and one Away-from-Camera Cell 43 selected from the grid image in Figure 3.2. For each cell, we also show an example of image patch and together with the extracted edge at specific orientation. The horizontal axes of the two plots represent the features described in Section 3.2.1.

Figure 3.5: Using the Mall dataset as a study case: the figures depict the weight contributions of neighbouring cells to cells 11, 51, and 55, which are highlighted using black boxes (refer Figure 3.2 for cell index). Red colour in the heat maps represents a higher weight contribution i.e. more information sharing.

3.3.4 Evaluation of Local Feature Mining of Our Model

The advantage of our multi-output framework for localised crowd counting is to mine local feature importance for supporting crowd density estimation. As Figure 3.4 shows, at different depths in a scene, certain types of features can play a more important role in estimating specific localised crowd density. Specifically, the plots in Figure 3.4 shows that the edge orientation 60° (mainly corresponds to shoulder edges) in the Away-from-Camera Cell 43 exhibited a higher importance as compared to other edge orientations; whilst the edge orientation 90° (mainly corresponds to

	Region 1(R1)		
	mae	mse	mde
MLR	0.82	1.45	0.3611
MORR	**0.76**	**1.22**	**0.3317**

	Region 2 (R2)		
	mae	mse	mde
MLR	0.71	1.24	0.3317
MORR	**0.67**	**1.12**	**0.3061**

	Scalability (seconds)	
	Time-tr	Time-te
MLR	17.274	0.1028
MORR	14.848	0.0196

Table 3.3: Localised counting performance on two busy localised regions in the Mall dataset. Region 1 consists of Cells 11, 12, 19, and 20, while Region 2 includes Cells 43, 44, 51, and 52. Time-tr and Time-te denote the training time and testing time respectively.

the torso edges) in the Close-to-Camera Cell 6 was automatically assigned a higher weight. To provide a closer look on the different edge orientations at different cells, we depicted two example image patches and the associated edges extracted from cells 6 and 43 at the bottom of Figure 3.4. The results suggest that the weights learned using our model matched our intuition, i.e. different features, e.g. shoulder and torso edges, would have different importance to estimation at different depths of a scene.

3.3.5 Evaluation of Information Sharing Among Regions

To demonstrate that spatially localised regions in a scene can indeed share information with each other, we used the Mall dataset as a study case and selected three cells to profile how other neighbouring cells contributed to their count estimation. The degree of information sharing (or evidence support) can be quantified by summing the absolute weights of a neighbouring cell that contribute to the count of a cell that we are interested in. We repeated this step for all neighbouring cells in the image space. The weight contribution/information sharing can be transformed into a heat map as shown in Figure 3.5. Evidently, the closer the neighbouring cells to the selected cell region, the more information were shared. This observation suggests that our model is capable of seeking evidence support from other cell regions to achieve a more accurate counting estimate.

3.3.6 Comparison With Multiple Localised Regression Model

As shown in Table 3.2, our multi-output regression model outperformed the multiple local regression model (MLR). It is interesting to note that the performance of the MLR is even worse than the two global regression models, although it was motivated to overcome the limitations of global regression models [180]. Since the MLR model also measures the importance of different

features in different local regions as our model, this result highlights the importance of exploiting the correlation between features across regions and sharing information across regions to achieve more robust crowd counting. Without this information sharing, achieved by the multiple structure output regression model formulated in this chapter, the local measures are too noisy and brittle to be relied upon in isolation for estimating density. Importantly, our single model based regression approach is more computationally scalable compared to the MLR model, e.g. compared to MLR, MORR is 3-5 times faster in both training and testing. More details about training and testing time for MLR and MORR are given in Table 3.3.

3.3.7 Analysis of Localised Counting Accuracy

To demonstrate the effectiveness of the proposed MORR in localised counting, we selected two busy regions (right in front of two shops) in the Mall dataset and compared the performance of MORR against MLR. The selected regions are depicted in a figure together with the results on localised crowd counting in Table 3.3. As compared to MLR, our MORR achieved more accurate localised counting, and yet faster training and testing time. The results again suggest the importance of information sharing among regions.

3.4 Summary

We presented a single multi-output regression model capable of spatially localised crowd counting. Instead of building multiple localised regressors as adopted by existing techniques, our approach utilises a single joint regressor taking concatenated multiple localised imagery features as input for learning spatially localised crowd counts as multi-outputs. Our model outperforms multiple localised regressors on a challenging shopping mall dataset owing to its inbuilt ability for feature mining according to changing crowd conditions presented in different local spatial cell regions in the scene. On the other hand, it also compares favourably against existing single global regressor based crowd counting models. Extensive and comparative experimental results demonstrate the effectiveness of our method. On balance, by exploiting the shared characteristics of spatial localised regions, a more accurate prediction of crowd density estimation can thus be obtained. However, latent dependent correlation in feature space is mined spatially, which limits the framework to such a specific problem (i.e. crowd density estimation). Evidently, the common characteristics of label space in regression are shared with cumulative dependent nature, which

has not been exploited in previous literatures. In the light of this, to exploit a framework can capture the cumulative dependent nature of labels in regression is a significant topic, which will be presented in the next chapter.

Chapter 4

From Crowd Density Estimation to Age Estimation: Cumulative Attribute Space

---

A number of computer vision problems concern the estimation of a scalar value given a high dimensional feature input vector. Examples of such problems include age estimation from facial images [51,53,64,65,182,194], crowd counting [19,23,30,117], and human body/face pose (view angle) estimation [63, 127, 183]. Such a scalar value can vary continuously within a certain range but is often assumed to be discrete (e.g. human age and people count), and its estimation can be obtained by solving a multi-class classification problem [60, 94]. Such a multi-class labelling treatment of scalar value estimation assumes implicitly that each scalar output value (a label) is independent from other possible values (labels). On the contrary, human age and people-count are strongly correlated and neighbouring values have closer similarities than those further apart, e.g. a human face of 50 years old is more similar to that of 49 than that of 10. To exploit this observation, most existing approaches to the problem consider a regression solution in which a mapping function is learned explicitly between high dimensional feature input vectors and scalar output values [19, 23, 30, 51, 53, 64, 65, 182, 194]. However, there are two major challenges for learning a good regression function for solving such a problem: (1) inconsistent and incomplete features, (2) sparse and imbalanced training data.

In general, regression based interpretation suffers from large feature variations caused by both viewing conditions and visual inconsistency in interpretation. For instance, people of the same age can appear visually very different, e.g. the images were taken under very different

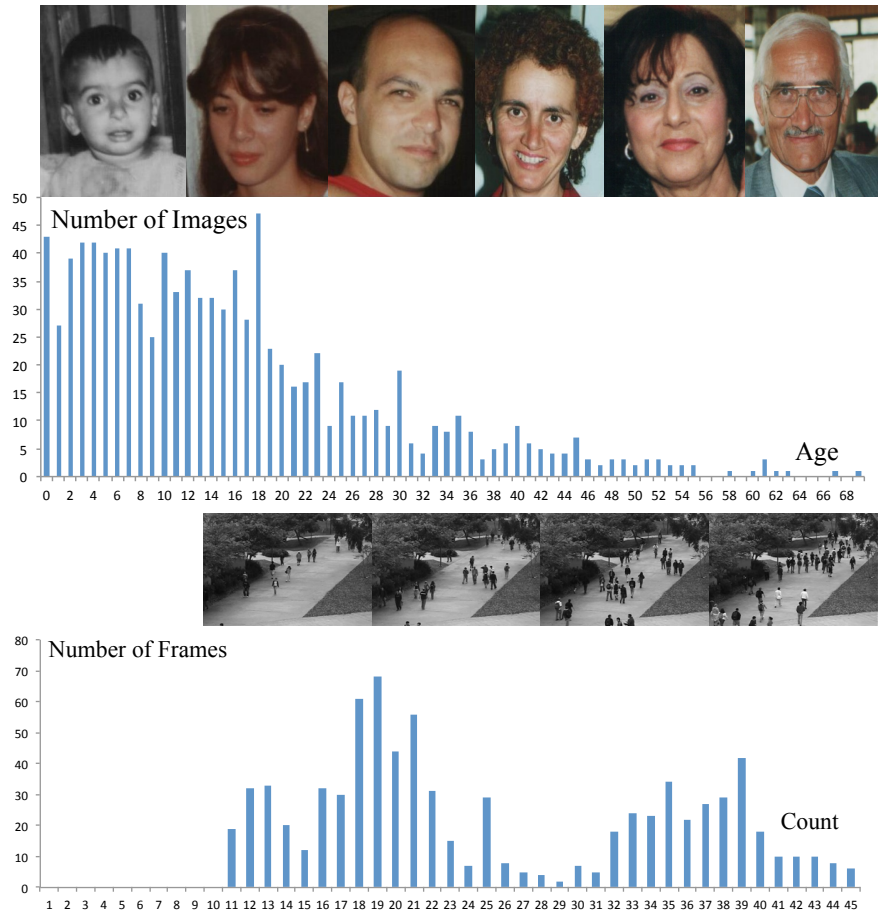


Figure 4.1: Age estimation and crowd counting both suffer from sparse and imbalanced training data distribution. Top: FG-NET facial age dataset. Bottom: UCSD crowd dataset.

lighting conditions (extrinsic condition change) or images of very different people of the same age (intrinsic condition change). In addition to lighting and viewing angles, occlusion can also cause crowd frames of the same people-count to appear significantly different. Existing regression techniques have mostly focused on addressing the challenge of feature inconsistency by constructing a low-level feature representation robust against both the intrinsic and extrinsic condition changes [65, 183]. There are less efforts on addressing the second challenge on sparse and imbalanced data.

Accurately labelled facial images for human age estimation and public space video data for crowd counting are generally sparse and imbalanced due to inherent ambiguities in annotation and a lack of sufficient samples for covering the data distribution. For example, despite large quantities of facial images available publically, e.g. from Flickr, annotating the true age of a facial image can be very unreliable [51,130]. As a result, benchmarking datasets such as FG-NET [25,60,64,194] and MORPH [25,60] contain very limited samples of each age group and consist of faces of true ages rather than annotated age. Figure 4.1 shows that in the FG-NET dataset, at

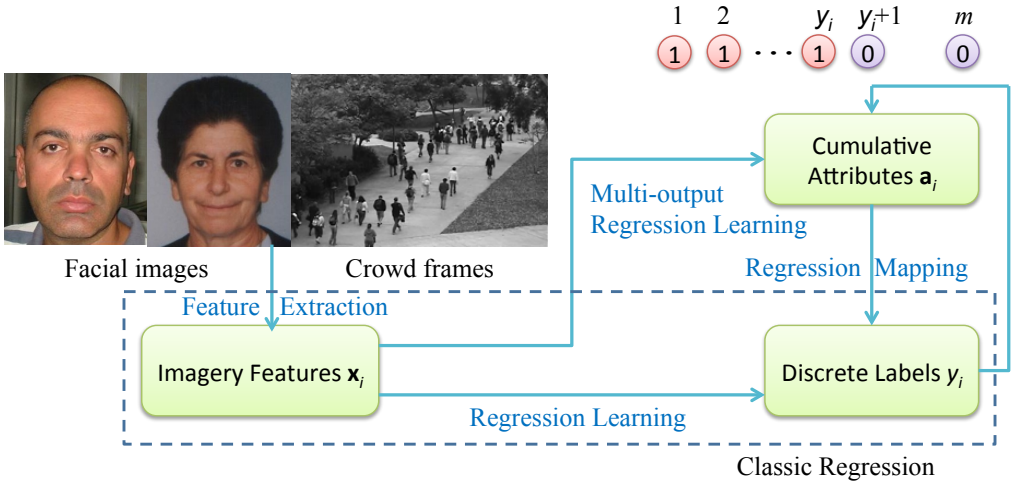


Figure 4.2: The pipeline of our framework compared with conventional regression framework.

most 46 images are available for each age group and the distribution is highly imbalanced across the age groups. This is rather sparse given that the faces belong to different genders and ethnical groups (therefore compounded by inconsistent visual features). Even though annotating crowd images can be made more reliable, annotating people count exhaustively for all possible values is laborious and often practically infeasible, e.g. a public place as shown in Figure 4.1 may never exhibit fewer than 10 people or greater than 50 people in any realistic time gap. Consequently existing crowd benchmarking datasets such as UCSD [19, 23, 30] are also sparse. Moreover, the sparseness in training data also implies that there are often gaps in training samples where no imagery sample is available for mapping onto certain output values causing difficulties in learning the regression mapping function.

In this work, we consider that the two challenges above are related in the sense that the feature inconsistency problem is compounded by sparse and imbalanced training data and vice versa, and they need be tackled jointly in modelling and explicitly in representation. To that end, we propose a novel cumulative attribute based representation for learning a regression model. Attributes have been successfully applied for solving various computer vision problems by classification [52, 93, 96, 110], but have never been used for regression to the best of our knowledge. Attribute models are designed to solve the data sparsity problem by exploiting shared characteristics between different classes. These common characteristics are either defined manually by human a priori knowledge [93, 96] or discovered automatically from data [52, 110]. Existing attribute learning methods cannot be directly applied to our regression problem because: (1) Attributes need to be discriminative to be useful. For classification, it is natural to identify dis-

criminative attributes for differentiating classes. Discriminative attributes can also be discovered by learning a discriminative model [110]. However, for learning a regression model it is much less clear what is discriminative and more importantly what can be shared across different scalar output values when those values change continuously. (2) Existing attribute definitions do not reflect nor exploit the unique characteristic of neighbouring scalar output values sharing more similarities than those further apart.

4.1 The Concept

Our notion of cumulative attributes aims to explore the spirit of the conventional discriminative attribute for addressing sparse training data, whilst is specifically designed for addressing the regression problem. More specifically, each attribute is not only discriminative but also cumulative in constraining all other attribute values depending on its relative positioning in value: each attribute separates all training images into two groups (binary) by a label (e.g. an age). For instance, for learning a regression model for age estimation, if there are 70 age groups, there will be 69 binary attributes, each separating facial images above a certain age from all those below. By cumulative attributes, we consider each attribute cumulatively conditioning all other attributes. That is, for a person of 50, not only the corresponding attribute 50 is positive, but also from 1 all the way to 49 are conditionally positive. This is designed specifically to capture the unique correlation of data samples so that those with neighbouring scalar output values share more than those further away in our cumulative attribute space. Critically, this cumulative nature is also able to cope with sparse and imbalanced data distribution more effectively. In particular, by utilising all data samples for discriminating each attribute regardless the availability of labelled data for that attribute (value) alone, sparsity problem is mitigated. The cumulative nature of the attribute also greatly reduce the ill-effects of imbalanced data, e.g. even if there was no sample for a certain age value (attribute), that attribute is positively assigned by any samples of lower age than the considered value, thus can be learned indirectly using plenty of neighbouring samples.

The pipeline of our framework is illustrated in Figure 4.2. Once cumulative attributes are constructed from the scalar values of training samples, a two-layer regression framework is employed. Firstly, given any low-level feature presentation of the image, we learn a multi-output regression model to map the feature inputs to an intermediate attribute space. To that end, a single structured output model is learned to correlate explicitly different attributes. Secondly, an-

other regression model is learned to estimate the scalar output using the attribute representation as input. Extensive experiments are carried out using benchmarking age estimation and crowd counting datasets and show that (1) our cumulative attribute representation improves generally the age estimation and crowd counting accuracy over the state-of-the-art with standard image feature representations, (2) the improvement is particularly significant when the training data is sparse and imbalanced.

4.2 Methodology

As shown in Figure 4.2, our cumulative attributes can be considered as an intermediate-level semantic representation that bridges the gap between any low-level features and a regression model given sparse annotation. During training our cumulative attribute based regression framework consists of the following steps:

1. Given a set of training images, we extract low-level imagery features and the scalar output value (e.g. age or people count) is converted into a binary cumulative attribute vector (Section 4.2.1).

2. A cumulative attribute representation is computed so that given an image, its cumulative attributes can be assigned and used as an intermediate representation of the image. Specifically, a single multi-output regression model is learned to evaluate and assign all attributes simultaneously (Section 4.2.2).

3. A second layer single output regression model is learned to map the attribute representation to the scalar output value (Section 4.2.3).

During testing, given an unseen image, the cumulative attribute vector is first computed using the multi-output regression model with the low-level imagery features as input. The cumulative attribute vector is then fed into the single output regression model to estimate the scalar output value.

4.2.1 Cumulative Attribute

Given a training image/frame i, where $i = 1, 2 \ldots \mathrm{N}$ and N denotes the total number of training images/frames, we firstly extract low-level imagery features $\mathbf{x}_i$ from the whole image/frame. This can be Active Appearance Model features [35] for age estimation and foreground & edges

& GLCM features [19,30] for crowd counting. Any other features in the literature can be equally applied. Secondly, normalization on the feature data including scale normalization and extra perspective normalization [19] for crowd counting are carried out.

Now for the ith training data point, the known scalar value y_i (e.g. age and people count) is converted into a cumulative attribute vector $\mathbf{a}_i$. The dimensionality of the vector $\mathbf{a}_i$, denoted as m, depends on the value range of y. Typically, for age or crowd count, there is an upper limit, e.g. 70 for a certain age dataset and 100 for a certain crowd scene. This upper limit will be used as the value of m. Formally, given N training samples $\{(\mathbf{x},y)\}_i, i = 1,2\ldots\text{N}$, the jth element of the cumulative attribute vector for the ith sample assumes a binary value:

$$a_i^j = \begin{cases} 1, & \text{when } \; j \leqslant y_i, \\ 0, & \text{when } \; j > y_i, \end{cases}$$

where $j = 1,2,\ldots,m$. Evidently, for the ith attribute vector $\mathbf{a}_i$, the first y_i attribute elements are all "ones" and the rest $m - y_i$ elements are all "zeros".

In comparison , a non-cumulative attribute (NCA) is constructed as follows:

$$a_i^j = \begin{cases} 1, & \text{when } \; j = y_i, \\ 0, & \text{when } \; j \neq y_i. \end{cases}$$

Note, only one element of a non-cumulative attribute vector $\mathbf{a}_i$ is one and all the remaining elements are zero. There is thus a critical difference between our CA representation and the conventional NCA representation: with the CA representation, data points with neighbouring scalar values are represented by a very similar attribute set, whilst with conventional NCA representations, the difference between the attributes of two data points of any scalar value is the same. For example, a face of age 40 and another face of age 41 represented using a 69D CA vector will have only one element that is different, whilst the number of different attribute elements increases to 30 for a face of age 10. On the other hand, using a NCA representation, there is always a single element difference no matter how different the ages are and how the two faces look alike. Our cumulative attributes thus capture a better representation of a continuously changing value for object appearance, corresponding directly to a scalar output value change continuously for learning a regression function. Our experiments in Section 4.3.3 show the distinct advantages of

using CA over NCA for both age estimation and crowd counting.

4.2.2 Joint Attribute Learning

Now the training set is represented as $\{(\mathbf{x},\mathbf{a},y)\}_i, i = 1,2\ldots\mathrm{N}$. We need to learn the mapping relationships between both $\mathbf{x}$ and $\mathbf{a}$, and $\mathbf{a}$ and y. In this section we focus on the former. Most existing attribute learning methods aim to establish a mapping between $\mathbf{x}$ and each element of $\mathbf{a}$ independently using a binary classifier such as a support vector machine. However, this is not only making the false assumption that different attributes are independent from each other, but also computationally expensive. In our work, we estimate the mappings of all m attributes simultaneously by learning a multi-output regression function, in particular, a multivariate ridge regression function [5, 67]. In its conventional form, a ridge regression function learns a single output mapping. Recently, multivariate ridge regression [5, 30] has been exploited for simultaneous output estimation. Following established design principle of multi-task learning [7, 8, 85, 152], we formulate the following multi-output attribute learning problem. Given $\mathbf{x}_i$ and a_i^j being low-level features of the ith image and the jth element of its corresponding attribute vector, the objective function for the jth attribute is written as:

$$\min \quad \left(\frac{1}{2}\|\mathbf{w}^j\|_2^2 + C\sum_{i=1}^{\mathrm{N}} \mathrm{loss}(a_i^j, f^j(\mathbf{x}_i))\right),$$

where $f^j(\mathbf{u}) = \mathbf{w}^j\mathbf{u} + b^j$ and $\mathrm{loss}(\cdot)$ denotes the loss function. Hence, a joint attribute learning by multi-output regression is formulated as

$$\min \quad \sum_{j=1}^{m}(\frac{1}{2}\|\mathbf{w}^j\|_2^2 + C\sum_{i=1}^{\mathrm{N}} \mathrm{loss}(a_i^j, f^j(\mathbf{x}_i))).$$

For simplifying the above without losing generality, the quadratic loss function is considered. The objective function of the joint attribute learning is then given as:

$$\min \quad \frac{1}{2}\|\mathbf{W}\|_F^2 + C\sum_{i=1}^{\mathrm{N}} \|\mathbf{a}_i^T - (\mathbf{x}_i^T\mathbf{W} + \mathbf{b})\|_F^2, \tag{4.1}$$

where $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, \ldots, \mathbf{w}^j, \ldots, \mathbf{w}^m]$ is the weight matrix, $\mathbf{a}_i = [a_i^1, a_i^2, \ldots, a_i^m]^T$ is the training attribute vector, and $\mathbf{b} = [b^1, b^2, \ldots, b^m]$ is the bias term. The model parameters $\mathbf{W}$ are estimated by solving an equality-constrained Quadratic Programming Problem, which has a closed-form

global optimal solution as follows:

$$\begin{bmatrix} \mathbf{W} \\ \mathbf{b} \end{bmatrix} = -(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T\mathbf{P},$$

where the positive semi-definite matrix $\mathbf{Q}$ and matrix $\mathbf{P}$ are given as

$$\mathbf{Q} = \begin{bmatrix} 2C\sum_{i=1}^{\mathrm{N}} \mathbf{x}_i\mathbf{x}_i^T + I & 2C\sum_{i=1}^{\mathrm{N}} \mathbf{x}_i \\ 2C\sum_{i=1}^{\mathrm{N}} \mathbf{x}_i^T & 2CN \end{bmatrix},$$

$$\mathbf{P} = \begin{bmatrix} -2C\sum_{i=1}^{\mathrm{N}} \mathbf{x}_i\mathbf{a}_i^T \\ -2C\sum_{i=1}^{\mathrm{N}} \mathbf{a}_i^T \end{bmatrix}.$$

The trade-off parameter C is determined by cross validation.

The weight matrix $\mathbf{W}$ plays an important role in transferring information between tasks thus modelling the correlation between different attributes. In particular, with the same feature representation, for each attribute $a_i^j, j = 1,2,\ldots,m$, we formulate our model to jointly weigh each attribute. In Equation (4.1), the jth column of matrix $\mathbf{W}$ is employed to weigh the imagery feature vector $\mathbf{x}_i$ for the jth binary attribute in corresponding attribute learning, i.e. the jth element of $\mathbf{a}_i$. Since the residual error of all attribute learning tasks are penalized jointly by the Frobenius-norm, this multi-output model can capture the correlation between different attributes explicitly.

4.2.3 Mapping Attributes to Scalar Output

To estimate the mapping between $\mathbf{a}$ and y, first the low-level feature $\mathbf{x}$ is mapped onto our cumulative attribute space using the learned multi-output regression model above. With each image now represented as $\hat{\mathbf{a}}_i \in \mathbb{R}^m$ and the corresponding label (ground truth) $y_i \in \mathbb{R}$, where $i = 1,2\ldots\mathrm{N}$, a second-layer output regression model is learned. Note, this regression model has a single scalar output and any existing regression models used in the literature for either age estimation or crowd counting can be readily applied.

4.3 Experiments

To demonstrate the motivation of this work, we conduct both experiments in the applications of facial age estimation and crowd density estimation, each of which are with two benchmark datasets.

4.3.1 Datasets & Settings

Datasets – For age estimation, two widely used benchmarking datasets FG-NET [25,60,64,194] and MORPH [25, 60, 141] were used. Both datasets are designed primarily for learning person-independent age estimator and contain people of different ethnical origins. For crowd counting, experiments were conducted on the benchmarking UCSD [19, 23, 30] and the Mall [30] datasets which feature an outdoor and an indoor scene respectively. Details in Table 4.1 show that among the four datasets, FG-NET is the most sparse in terms of the average number of samples per scalar output value (MORPH is 5 times more densely sampled).

Data	$\mathbf{N_{i/f}}$	**R**
FG-NET [60]	1002	0–69
MORPH [141]	5475	16–77
UCSD [19]	2000	11–46
Mall [30]	2000	13–53

Table 4.1: Dataset details: $\mathbf{N_{i/f}}$ = number of images/frames, **R** = range of scalar output value.

Features – For age estimation, the low level image features are Active Appearance Model features [35], which is also reviewed in Section 2.2.5 with more details. This feature representation is widely used in recent approaches [25,60,64,181,182,194]. For crowd counting, three types of image features, i.e. foreground segments, edge features, and local texture features, are adopted as in [19,30]. Note that, to use these features, all frames of crowd databases were transformed to gray-scale prior to feature extraction.

Settings – For FG-NET, we followed the same leave-one-person-out setting as in [25,64,181,182, 194]. For MORPH we randomly split the dataset into 80% training data and the rest 20% testing data and repeated the experiments 30 times as in [25]. For crowd counting, we followed the same training and testing partitions as in [30], i.e. we employed Frames $601-1400$ in UCSD dataset and Frames $1-800$ in Mall dataset respectively for training, while the rest frames were used for testing. For the single output regression model (Section 4.2.3), Support Vector Regression (SVR) with RBF kernel and Ridge Regression (RR) were employed for age estimation and crowd counting respectively, owing to their strong performance reported in the literature for age [64,65] and crowd [30] respectively. However, any regression models can be used.

Evaluation Metrics – For age estimation, we employed two evaluation metrics, namely *mean*

absolute error (mae) defined in the previous chapter and *cumulative score* (cs),

$$\varepsilon_{\text{cs}} = (M_{abs<L}/M) \times 100\%,$$

which $M_{abs<L}$ and M denote the number of testing images with absolute error less than error level L and the total number of testing images [60], and we set the same error level $L = 5$ as in [25]. Three metrics employed in [30], namely *mean absolute error* (mae), *mean squared error* (mse), and *mean deviation error* (mde) were employed for evaluating the performance of crowd counting, which are defined in Chapter 3. Among all five metrics, only for cs higher value means better performance.

4.3.2 Comparison with State-of-the-Arts

Method	FG-NET [60]		MORPH [141]	
	mae	cs	mae	cs
AGES [60]	6.77	–	8.83	–
RUN [182]	5.78	–	–	–
Ranking [181]	5.33	–	–	–
RED-SVM [24]	5.24	–	6.49	–
LARR [64]	5.07	–	–	–
MTWGP [194]	4.83	–	6.28	–
OHRank [25]	4.85	74.4%	**5.69**	56.3%
SVR [64]	5.66	68.0%	5.77	57.1%
CA-SVR	**4.67**	**74.5**%	5.88	**57.9**%

Table 4.2: Age estimation performance comparison.

Age estimation – Our model (CA-SVR) is compared with a number of contemporary published results in Table 4.2. Most of the methods compared are regression based except AGES [60], RED-SVM [24] and OHRank [25], and use the same AAM features except AGES [60]. For FG-NET dataset, our model obtained the best results so far on both mae and cs metrics. Note that compared with SVR [64], identical low level feature and single output regression models were used. The only difference is in the input to the regression model: low level feature directly for SVR and our cumulative attributes for CA-SVR. This change of representation brings an significant improvement (17.5% decrease in mde and 9.6% relative increase in cs). The best performance reported so far on FG-NET is the Ordinal Hyperplane Rank model (OHRank) [25]. OHRank can also cope with the sparse data problem, but the rankers are not cumulative therefore do not share mutual information, and they do not benefit from an intermediate representation. As shown in previous section, it is in the order of four magnitudes slower than our model in model

training[1]. On the MORPH dataset, our CA-SVR gives comparable result to the best reported so far (OHRank) on mae, but best performance measured by cs. As the key difference between the FG-NET and MORPH dataset is data sparsity and the number of age groups without samples, it is evident from these results that the advantage of our cumulative attribute based regression model is more significant given sparse and imbalanced data. This is further supported by our missing data experiments reported in Section 4.3.4.

Method	UCSD [19]			Mall [30]		
	mae	mse	mde	mae	mse	mde
LSSVR [159]	2.20	7.29	0.107	3.51	18.2	0.108
KRR [5]	2.16	7.45	0.107	3.51	18.1	0.108
RFR [106]	2.42	8.47	0.116	3.91	21.5	0.121
GPR [19]	2.24	7.97	0.112	3.72	20.1	0.115
RR [30]	2.25	7.82	0.110	3.59	19.0	0.110
CA-RR	**2.07**	**6.86**	**0.102**	**3.43**	**17.7**	**0.105**

Table 4.3: Crowd counting performance comparison.

Crowd counting – Table 4.3 compares crowd estimation performances of six different methods, all based on regression, using the two benchmarking datasets. The result shows that the cumulative attribute based model (CA-RR) performs the best for both datasets and using all three metrics. The most direct effect of using our cumulative attribute representation can be seen by comparing RR [30] with CA-RR. CA-RR clearly outperforms RR using all three measures. Since both have the same low level feature input and use the same single output regression model, the performance gain can only be explained by the superior representation by our cumulative attribute space. Improved performance can also been seen by comparing CA-RR with a number of recently proposed models [5, 19, 106, 159], all of which use the same features as input and differ only in the regression model used.

4.3.3 Cumulative vs. Non-Cumulative Attributes

Methods	FG-NET [60]		MORPH [141]	
	mae	cs	mae	cs
NCA-SVR	8.95	41.8%	7.28	44.2%
CA-SVR	**4.67**	**74.5**%	**5.88**	**57.9**%

Table 4.4: Cumulative vs. non-cumulative attributes on age estimation.

A key novelty of our model is the cumulative attribute representation. As explained in Sec-

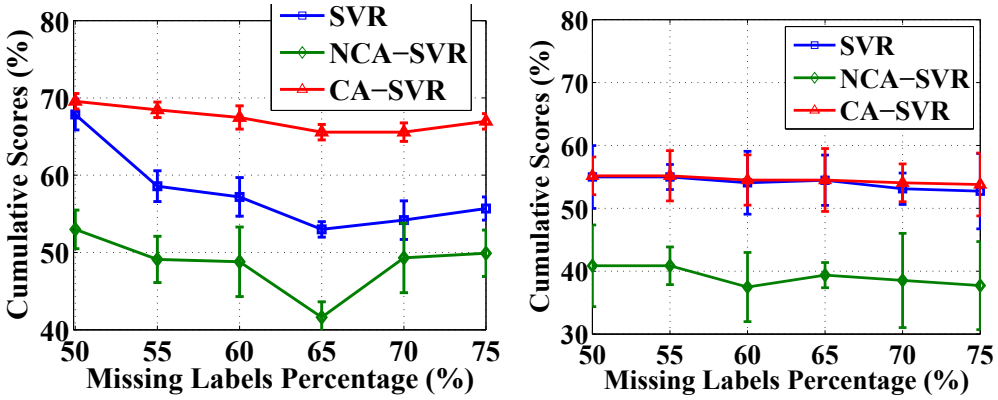
[1]The results of OHRank were based on our implementation and are slightly lower than those reported in [25] with FG-NET dataset and slightly higher with MORPH dataset.

Methods	UCSD [19]			Mall [30]		
	mae	mse	mde	mae	mse	mde
NCA-RR	2.85	11.9	0.137	4.31	25.8	0.131
CA-RR	**2.07**	**6.86**	**0.102**	**3.43**	**17.7**	**0.105**

Table 4.5: Cumulative vs. non-cumulative attributes on crowd counting.

tion 4.2.1, compared with the conventional non-cumulative (NCA) attributes, the unique characteristics of our cumulative attributes (CA) is that data points of neighbouring scalar value are designed to be close to each other in the attribute space. It is evident from Tables 4.4 and 4.5 that constructing such cumulative attributes is a significant advantage for a regression model that performs age estimation and crowd counting.

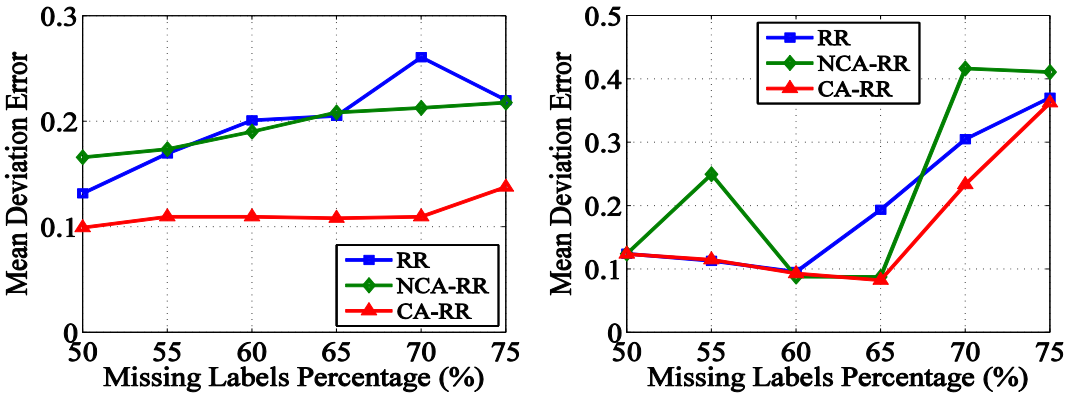
4.3.4 Against Sparse and Imbalanced Data



(a) FG-NET (b) MORPH

Figure 4.3: Age estimation performance with sparse and imbalanced data measured using cumulative scores (the higher the better). To illustrate the stability of attribute-based model (CA-SVR) and non-attribute based models (SVR and NCA-SVR), the deviation of performance metrics in the form of error bars are also added here.

Figures 4.3 and 4.4 evaluate our model when the training data become more and more sparse and imbalanced. Data of certain age groups and certain crowd counts were removed to make the data more sparse and imbalanced. For age estimation, since the two dataset have few missing age groups, we randomly selected a fixed number of age groups, each time to remove and then train the model. For the crowd counting dataset, this way of removing data would be less effective because the mapping between the low level features and the scalar count numbers is more linear. Therefore, a different strategy for removing samples is adopted. That is, we start from the middle of count number ($26-30$ for missing 10% count groups in our case) and then



(a) UCSD (b) Mall

Figure 4.4: Crowd counting performance measured by mean deviation error (the lower the better).

remove an entire chunk of count groups. It is evident from Figures 4.3 and 4.4 when more training data were removed, the performance of all the models degrades. However, our model's performance degraded more gracefully, resulting in the bigger performance gain over both the non-attribute based models (SVR and RR for age and crowd respectively) and non-cumulative attribute methods. Moreover, in age estimation, for the repeated experiments about removing randomly-selected labels, our model can also achieve superior performance with less diviation than non-attribute based models. These results further validate our early observation that the construction of a cumulative attribute space is uniquely effective for coping with sparse and imbalanced training data, a common problem in learning regression functions.

4.3.5 Learning Attributes Jointly vs. Independently

Methods	FG-NET [60]		UCSD [19]		
Original Dataset	mae	cs	mae	mse	mde
i-CA	4.73	73.7%	**2.07**	7.09	**0.102**
j-CA	**4.67**	**74.5%**	**2.07**	**6.86**	**0.102**
Missing 75% labels	mae	cs	mae	mse	mde
i-CA	6.45	55.6%	2.87	13.3	0.139
j-CA	**5.51**	**66.9%**	**2.79**	**12.6**	**0.137**

Table 4.6: Jointly learning cumulative attributes (j-CA) vs. independently learning cumulative attributes (i-CA).

Instead of learning all attributes jointly using our multi-out regression model, experiments were conducted to learn each attribute independently using a single out ridge regression model. Table 4.6 shows that comparing with the jointly learned attributes, the independently learned

attributes led to poorer performance. In particular, for more imbalanced data with the removal of 75% labels from the original training dataset, our joint learning model yields more significant advantage on both the FG-NET age dataset and the UCSD crowd dataset. This is because that for sparse data, information sharing between attributes can contribute to improve robustness because of jointly penalizing the errors in different attributes.

4.3.6 Computational Cost

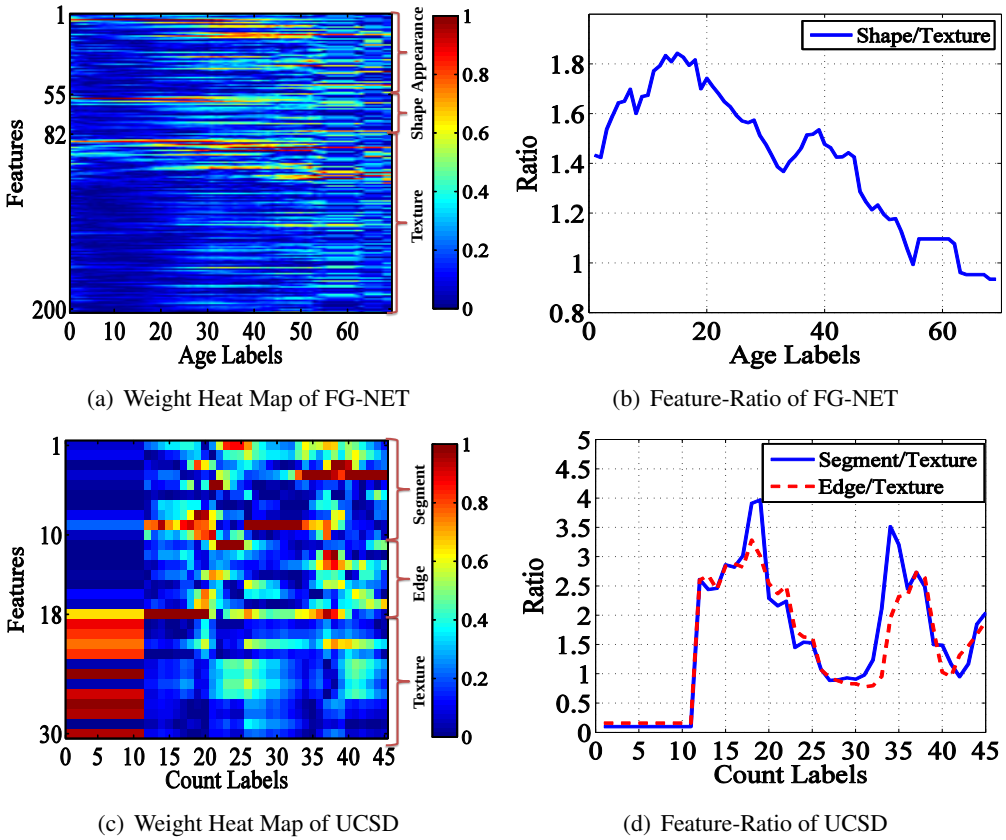
Methods	Age (mins)		Crowd (secs)	
	FG-NET [60]	MORPH [141]	UCSD [19]	Mall [30]
OHRank	1.30×10^4	3.02×10^4	–	–
SVR [64]	2.69×10^0	2.08×10^1	–	–
RR [30]	–	–	**0.70**	**0.67**
CA	$\mathbf{8.91 \times 10^{-1}}$	$\mathbf{6.10 \times 10^0}$	1.57	1.52

Table 4.7: Model training time required by different models.

Table 4.7 shows the training time for four different models. It is evident that the proposed cumulative attribute based model is extremely fast to learn owing to its closed form solution based on a multi-output regression model (see Section 4.2.2). For age estimation, it is even faster to train than the non-attribute based model with the same single output regression. The closest competitor for age estimation accuracy, OHRank [25] is four orders of magnitude (10^4) slower than our model (under 7 mins). This is because after mapping the low level image features to the cumulative attribute space, dimensionality reduction is achieved as a by-product resulting faster single output regression model training. For crowd counting, RR [30] is faster than CA. This is because the cumulative attribute space has a similar dimension as the original low-level feature and CA has the additional step of estimating the attribute values. Nevertheless, both are very fast to train (under 2 sec).

4.3.7 What is Learned by Cumulative Attributes?

To answer this question, Figures 4.5(a) and (c) visualise the weight matrix $\mathbf{W}$ in Formulation (4.1) which shows how different low level features are weighted for different scalar value groups. For age estimation, the AAM features capture the shape and texture characteristics of a human face. It is known [51] that at earlier ages, the human aging process is mainly reflected by the facial bone change (getting mature) resulting in shape changes. Entering adulthood, texture change gradually starts to play a more important role because aging is now more concerned with skin



(a) Weight Heat Map of FG-NET

(b) Feature-Ratio of FG-NET

(c) Weight Heat Map of UCSD

(d) Feature-Ratio of UCSD

Figure 4.5: Visualization of the importance of different features for cumulative attributes. Weights of each type of features were averaged for computing the weight ratio between different types of features.

changes (e.g. having more wrinkles). Figures 4.5(a) and (b) show that our learned cumulative attribute indeed capture this phenomenon rather well. In particular, the shape features are the most important ones that separate attributes correspond to young ages (< 20), while texture features become more and more important for elder ages. For crowd counting, the 30 low level features contain foreground segment area, edge features and texture features. Segment and edge features would in general be more sensitive to the different crowdedness levels compared to the texture feature. That is, more people in the scene normally means larger foreground regions and more edges. This is also reflected by the learned weights shown in Figures 4.5(c) and (d).

4.4 Summary

We have introduced a novel cumulative attribute based framework for solving a number of computer vision problems invoking the need for regression estimation. Noisy and sparse low level visual features are mapped onto a cumulative attribute space where each dimension is designed

specifically to give a clear semantic meaning that captures how the scalar output (e.g. age, people count) changes continuously. It requires no additional human annotation to assign attributes and can be estimated efficiently and robustly given sparse and imbalanced training data. Extensive experiments show the effectiveness and efficiency of the proposed model for both age estimation and crowd counting. This advantage of our approach is particularly significant when the training data is sparse and imbalanced. On the other hand, our proposed cumulative attribute concept is to capture the scalar-formed regression labels in the sense that the neighbouring output values shares more than those far away from in cumulative attribute space. However, when extending single-output to multi-output (i.e. from scalar-valued to vector-formed), cumulative attribute concept may not work well because of the missing latent correlation between output variables. To capture the latent structure of multi-label variables, structured learning for regression can be a feasible solution. In next chapter, we will introduce structural support vector machines for estimating 2D human posture.

Chapter 5

Structural Output Regression Learning for Human Pose Estimation

---

The problem of estimating the configuration of a person's body parts have attracted increasing attention from computer vision researchers. Human body pose estimation has been widely used in many applications such as video surveillance [73], human-computer interface [129] and computer games [124]. However, despite the best efforts in the past decades, the human pose estimation problem estimation, especially under cluttered and uncontrolled environments, remains unsolved due to the ambiguity caused by self-occlusion, body configuration and low contrast between foreground and background.

Most existing works on human pose estimation focus on model-based methods, which specify a rough approximation of the skeleton and then use such a model in conjunction with image measurements to estimate the best-fitting pose. Those model-based techniques are characterised by a kinematic model that relates constraints between body parts including kinematic constraints of articulated human as well as other constraints such as appearance constraints [44, 138]. Pictorial Structure Model was proposed by Felzenswalb *et al* [48], which uses a prior model to measure the likelihood of the location of each limb by using appearance terms. In [138], an image parsing method based on Pictorial Structure Model employs a priori human model representing the subject and updating the model continuously with edge and colour information of still images. Recently, a method based on progressively reducing the search space by employing image parsing has been proposed which achieves superior results [49]. Based on such an image-

specific color model, Eichner and Ferrari [44] use an enhanced pictorial method containing an appearance model describing hidden relationship between each body parts according to location priori within the foreground. Johnson and Everingham [80] try to add coherent appearance properties of each body parts to a Pictorial Structure Model in order to improve the results. In [81], clustering is performed to discover pose groupings in a pose space. This model is still based on the pictorial structure, thus is still a generative method. Despite its popularity, it is noted that using a generative model for estimating human pose may have some drawbacks, including a) not suitable for real-time application due to its slow inference algorithm and b) prone to over fitting given limited training data.

To overcome the drawbacks of generative methods, discriminative regression methods can be considered for human pose estimation which once trained can run very fast during testing. However, general discriminative regression methods such as Support Vector Regression (SVR) and multivariate ridge regression could only estimated the output pose parameters individually instead of in a global and structured manner. In detail, multivariate regression technique, specifically multivariate ridge regression, is proven to solve the localised crowd counting effectively and efficiently in the last chapter owing to mining the spatial correlation between local feature from localised subregions. In such a multi-input-multi-output (MIMO) framework, important correlated information between output variables are missing. In other words, those non-structured regression methods ignore the important information about the relevance between each body parts in our case.

5.1 The Concept

We consider that taking consideration of the important correlation between output variables can improve the estimation accuracy of performance. In this chapter, recently proposed structural support vector machines (i.e., Structural Support Vector Regression (SSVR) [12, 76, 79] and Latent Structural Support Vector Regression (LSSVR) [190]) in multi-input-multi-output frameworks are adopted for regression learning in 2D human upper body gesture estimation. Different from aforementioned non-structural multi-output regression framework (MORR) in last chapter as well as non-structural single-output regression technique (Support Vector Regression), both of aforementioned structural methods are designed to capture the dependency on structured input and structured output. Extensive experiments using public benchmarking datasets have been

carried out to demonstrate that: i) During testing, our methods could run much faster than generative methods; they are thus more suitable for real-time application. ii) Our methods generate acceptable results when the size of training database is reduced dramatically. iii) Compared to non-structured discriminative methods (e.g., Support Vector Regression and Multivariate Ridge Regression), structured methods achieve better performance owing to the ability to capture the important relevance information between outputs.

5.2 Methodology

In this section, we will present problem formulation and our methods in detail. For 2D human upper-body pose estimation in still images, we wish to find out the configuration of six human upper body parts (head, torso, and upper/lower right/left arms). For unconstrained still images, we know nothing about the person's appearance (e.g. what cloth she/he wears) and it is expensive to search the whole images. For effectively estimating human pose, one pre-processing step will be taken before model learning to reduce the possible space and improve the efficiency of our approach. In particular, we will use a pre-learned upper-body detector [44, 49] to localise the human body. We detect the upper body in each frame using a sliding window approach with a Histograms of Oriented Gradients representation of human appearance [38]. By using the upper-body detector, the search space will be reduced significantly. After the localization of the upper body, learned structured discriminative regression models are then used to estimate the body pose.

5.2.1 Model Input and Output

In unconstrained still images, low-contrast and diverse appearance could increase the difficulty of estimating human pose. It is therefore vital to extract informative appearance features as model input. In our model-free framework, for no kinematic model is used to constrain the estimation procedure, the features extracted from the detected upper body bounding should capture information that is useful for identifying different body parts and sensitive to body pose changes. To this end, Bag-of-word SIFT features are used as the input while the output for our regression models are structured coordinates. More specifically, for the i-th body part, the output are coordinates $[x_{1i}; y_{1i}; x_{2i}; y_{2i}]$. For using a pre-learned upper-body detector, multiple bounding boxes may exist in a single image. Note that for training, we will use ground truth location of upper bodies to

extract those features as model inputs. During testing, the body location is provided by the upper body detector. After localizing the upper body, we will extract the bag-of-words SIFT within the bounding boxes for model inputs. Randomly chosen descriptors are employed by K-means to generate a codebook with 400 clusters. In order to incorporate location information of each body parts into the model inputs, each bounding box is divided into $2 \times 2 = 4$ sub-regions. A 400 dimensional feature vector is then computed from each sub-region and the four feature vectors are concatenated into a 1600 dimensional feature vector as the final model input. The model output has a dimensionality of 24 (4 coordinates $\times$ 6 body parts).

5.2.2 Structural Support Vector Regression

The problem formulation for 2D upper-body pose estimation is as follows. For supervised learning, we have pairs of input and output x_i, y_i, where $i = 1, 2, \ldots N$ and N denotes the size of training set. x_i and y_i are feature vectors of 1600 and 24 dimensions respectively as described above. During training, each body part is mannually annotated and the value of y_i is computed from the annotation. The objective of model learning is learn a discriminative regression function as a linear combination of joint features [12, 76, 79, 163]:

$$\text{argmin}_y f_w(x,y) = w^T \Psi(x,y), \tag{5.1}$$

where w is a parameter vector and $\Psi(x,y)$ is a feature vector induced by a joint kernel $K(x,y,x',y') = \Psi(x,y)^T \Psi(x',y')$. The above structured Support Vector Regression problem is thus solved by estimating the parameter vector w. This can be formulated as the following optimisation problem:

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C\xi \\ s.t. \quad & \forall (\bar{y}_1, \ldots, \bar{y}_N) \in Y^N \\ & \frac{1}{N} w^T \sum_{j=1}^{N} [\Psi(x_j, y_j) - \Psi(x_j, \bar{y}_j)] \\ & \geq \frac{1}{N} \sum_{j=1}^{N} \Delta(y_j, \bar{y}_j) - \xi, \end{aligned} \tag{5.2}$$

where the loss function $\Delta(y_j, \bar{y}_j)$ and $\Psi(x,y)$ are problem-dependent. It is worth mentioning here that the above equation is a 1-slack formulation, which is more efficient than the original n-slack one. In our case, we will consider a square distance as the loss function for pose estimation.

According to the dual theory, we could easily get the dual problem for the above equation, which is used for constraint generation as well as the stability analysis [79]. It is worth pointing out that the above formulation for our problem could be margin-rescaling and there also have a slack-rescaling formulation, e.g., OP3 in [79], which is not effective for our problem. For human pose estimation, the kernel function which could induce the feature vector in Equation (5.2) is similar to joint RBF kernel [177]. That is,

$$K((x,y),(x',y')) = exp(-\|(x,y)-(x',y')\|^2).$$

For the output loss function we use the square difference of image feature vector. More specifically, the formulation is as follows

$$\begin{aligned}\Delta(y,y') &= \|\varphi(y)-\varphi(y')\|^2 \\ &= K(y,y)+K(y',y')-2K(y,y') \\ &= 2(1-K(y,y')).\end{aligned}$$

5.2.3 Latent Structural Support Vector Regression

In this subsection, Latent Structural Support Vector Regression is investigated and formulated. Latent Structural Support Vector Machine, was proposed by [79] to solve the classification problem. Here the model is extended so that it can be used for regression, i.e. the model outputs become continuous rather than discrete. The difference between Latent Structural Support Vector Regression and Structural Support Vector Regression is the introduction of latent variables in the model. With latent variables the model aims to capture not only the input-output relationship but also unobserved relationships, such as relationship between different body parts [47].

The detailed formulation using Latent Structural Support Vector Machine for our problem will be presented as the following. In comparison with the aforementioned Structural Support Vector Regression model, latent variable vector will be added to the joint feature vector $\Psi(x,y,h)$:

$$\text{argmin}_{(y,h)} f_w(x,y,h) = w^T\Psi(x,y,h). \tag{5.3}$$

Similar to the Structural Support Vector Regression presented in the last subsection, a joint kernel $K(x,y,h,x',y',h') = \Psi(x,y,h)^T\Psi(x',y',h')$ could be induced by $\Psi(x,y,h)$. As a result, the

optimisation problem of a Latent Structured Support Vector Regression model is written as:

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C \sum_{j=1}^{N} \xi_j \\ s.t. \quad & \forall (\bar{y}_1, \ldots, \bar{y}_N) \in Y^N \\ & w^T [\Psi(x_j, y_j, h) - \Psi(x_j, \bar{y}_j, \bar{h})] \geq \Delta(y_j, \bar{y}_j, \bar{h}) - \xi, \\ & \text{for} \quad j = 1, 2, \ldots N; \end{aligned} \tag{5.4}$$

where $x_j, y_j, j = 1, 2, \ldots N$ is the training pairs and $\bar{y}_j, j = 1, 2, \ldots N$ denotes prediction results approaching y_j during training procedure. Note that, for simplifying the formulation and increasing the efficiency, the loss function will not depend on $h_i^* = \text{argmin}_h w^T \Psi(x_i, y_i, h)$ but on the predicted latent variable $\bar{h}$ for practical application [190]. Evidently, Equation (5.4) could be reduced into the Structural Support Vector Regression formulation by removing the latent variables. In this work, the kernel function in the above formulation the same joint RBF function as Structural Support Vector Regression. In other words, the loss function only depends on input and structured output without latent variables. The optimization problem is solved using the Concave-Convex Procedure (CCCP) [190, 192], which is guaranteed to converge to a local minimum.

Before ending this section, we have one remark about three discriminative methods. Compared to Support Vector Regression, Structural Support Vector Regression and Latent Structural Support Vector Regression have the potential to solve more complicated regression problem owing to their ability to model structured outputs. However, the price to pay is the increased model complexity, which may imply higher computational cost. As a result, the tradeoff between complexity and accuracy needs to be determined according the application at hand and the amount of training data available. In particular, the latent variables adding into Structural Support Vector Regression means that more training data are required to learn the model in comparison with SSVR. In other words, when the training data size is small, LSSVR is more likely to suffer from model over-fitting resulting in worse performance.

5.3 Experiments

For the purpose of demonstrating the motivation of introducing structured output learning, i.e. to learn the inherent latent dependent relation between each element of output representation, the

following experiments are conducted.

5.3.1 Datasets and Settings

Experiments were carried out to demonstrate the effectiveness and efficiency of our models for human pose estimation. We used the same databases as in Ferrari et al.'s paper [44, 49] including cluttered images from the TV episodes *Buffy the Vampire Slayer* and highly challenging images from PASCAL VOC 2007 and 2008 datasets.

Three experiments were conducted, each of which differs in how the training/testing dataset were organised. In the first experiment, different sizes of randomly selected images from Buffy Episodes 3&4 and VOC 2007&2008 were employed as training sets, while the test sets were the same including 276 images from Buffy Episodes 2&5&6. In the second experiment, the training and test sets were replaced by Buffy Episodes 2&3&4&5&6 and Pascal 2007 containing 91 testing images respectively. In the third experiment, a more balanced training set was used which has the same number of different categories of poses selected from Buffy Episodes 3&4 and VOC 2007&2008, while the test set was the same as that used in the first experiment.

The inputs and outputs for the three models we compared (i.e., SVR, SSVR, and LSSVR) are the feature vectors and corresponding human body configuration. In order to increase the accuracy of estimation and speed up the computation, one pre-processing step was employed, that is, the upper-body detector [49], which will search the entire image to find the rough position and scale of people. It is evident that our results rely on the good performance of upper-body detector as pose estimation will only be performed in the detected regions. The detection rate of the bounding boxes detected was relatively high on the datasets we used. Specifically we achieved a detection accuracy of 0.8043 for the Buffy database in our testing sets. When there exit multiple detections in one image, multiple poses will estimated but only one pose will be selected for comparing with ground truth. This is because as the ground truth for each image of the Ferrari et al.'s database [44, 49] provides only one pose. Within each of the bounding boxes, 5000 bag-of-word SIFT features was extracted and we then used k-means to create a codebook consisting of 400 clusters. In all of our experiments, we use PCA [47] to reduce the dimension of the input image feature vector from the original 1600 (4 sub-regions with 400 histogram bins in each sub-region) to 20 dimensions in order to increase the efficiency. For SVR training, 24 Support Vector Regression will be trained independently, while for SSVR and LSSVR, the 24 outputs were estimated jointly. During LSSVR training, latent variables were manually labeled

according to different categorizes of poses (5 in our experiments). To evaluate the performance of different models, Percentage of Correctly estimated body Parts (PCP) will be used [44, 49], i.e., an estimated body part is deemed as correct if its segment endpoints lie within 50% of the length of the ground-truth segment from their annotated locations.

5.3.2 Computational Efficiency of the Proposed Models

The Experimental results are shown in Tables 5.1, 5.2 and 5.3 as well as Figures 5.1 and 5.2. Compared to the best results generated by generative models [44] (i.e., 78.1% obtained by using a training set of 1021 images and the same testing set as in our first experiment), our results are comparable. However, our discriminative models are much more efficient to compute. Specifically, using our discriminative methods, it took 5.86 seconds per image on average. On the other hand, using the generative method in [44] it took more than 70 seconds for testing one image. In other words, the testing time using the discriminative methods is more than 10 times faster in comparison with generative methods in [44,49]. Moreover, it is found that, by using less training images as Ferrari did [44], the structured methods (especially SSVR in all three experiments and LSSVR in the third experiments) which we propose for human pose estimation has a PCP not too much lower than that of the generative method.

5.3.3 Effect of Modelling Structured Output

Our results show that for all three experiments, the two structured regression models, particularly SSVR significantly outperform the regression model without modelling the structure of model outputs (SVR). This results show the importance of modelling output variable structure for the problem of human pose estimation. This is because the position of different body parts are typically highly correlated. Ignore the structure of them thus means that important information has been left unexplored.

5.3.4 Effect of Training Data Size

Tables 5.1, 5.2 and 5.3 shows that, with an increasing training set size, the performance of all three regression machines improves. Moreover, in all three experiments, SSVR shows the best generalisation capability for human pose estimation among the three discriminative methods. Figure **??** also shows the disparity in performance of the LSSVR over the three datasets. It is worth pointing out that LSSVR achieves the significantly worse result in the second experiments.

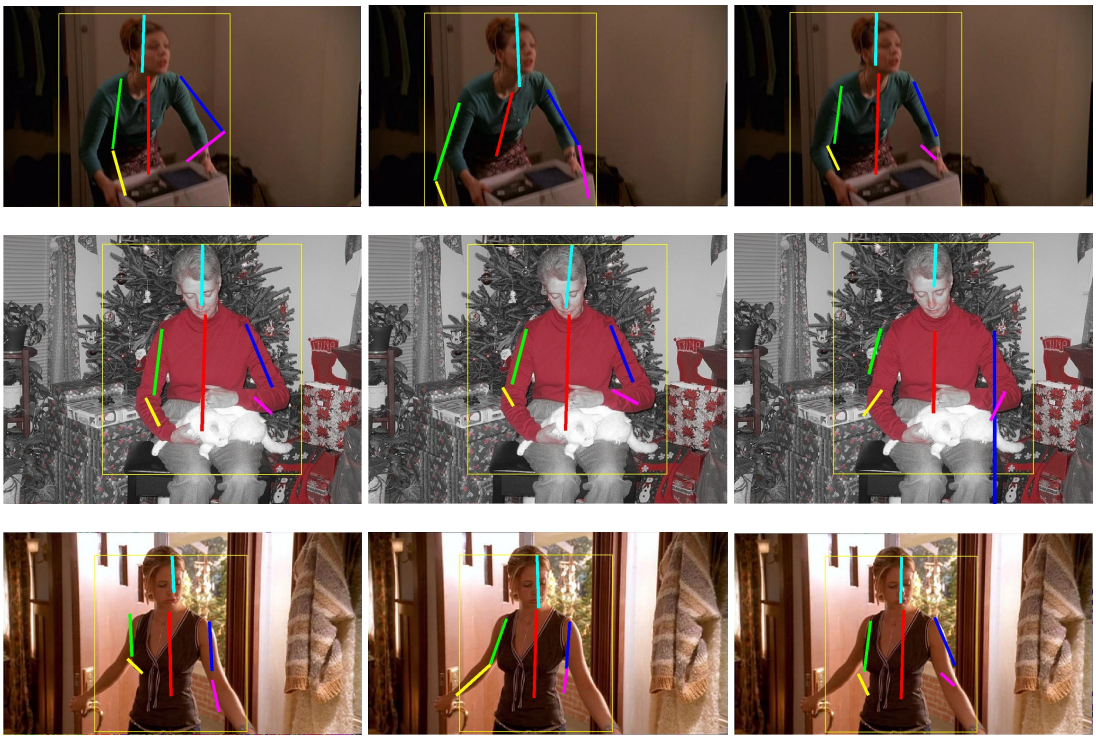


Figure 5.1: Illustrative results for testing Buffy and Pascal generated by LSSVR (Left), SSVR (Middle) and SVR (Right)

This is because in the Buffy datasets, most of the latent variables are the same (i.e. most of the poses in Buffy datasets are similar). In other words, the poor performance of LSSVR in the second experiment was due to the fact that the latent variables in the LSSVR model become redundant thus having a negative effect.

5.3.5 Effect of a Balanced Training Set

The third experiment was designed to demonstrate the importance of preparing a balanced training dataset when LSSVR is employed. Form Table 5.3, we can see that LSSVR outperforms SSVR and SVR when the size of training database is 400. In comparison, in our first experiment, LSSVR could only achieve superior performance to SSVR when the size of training database is 1021. This result indicates that when the training dataset is large enough and has the balanced number of different poses for each pose category, the performance generated by LSSVR can be superior to that of SSVR and SVR.

5.3.6 Multi-Output vs. Structural-Output Regression

In the definition, multi-output regression learning is based on the concept of joint learning, while structured output learning is to take the consideration of both features and other elements of out-

Figure 5.2: Illustrative results by our model-free discriminative methods with single detection for single person (Left), multiple detection for single person (Middle) and multiple detection for multiple detection for multiple person (Right). The top image of the right column shows that our method can estimate multiple poses for one image at the same time, while the middle and bottom images of the right column illustrate that multiple wrongly-estimated poses caused by over-enlarging bounding boxes for human upper body and localization.

Table 5.1: PCP with different size of randomly selected training datasets for the Buffy testing set (i.e., 276 images of Buffy Episodes 2, 5 and 6), where SoD denotes the size of training database.

SoD	50	100	200	400	1021
SVR	29.71%	29.79%	32.35%	40.93%	58.33%
SSVR	48.72%	51.21%	58.97%	67.49%	72.82%
LSSVR	42.88%	44.39%	47.47%	56.79%	73.79%

put simultaneously. Evidently, benefiting from exploiting the correlation between each output entry in the framework, structured output learning can outperform multi-output learning models in case that each element of output are dependent. For verifying the advantages of structural learning models, the following experiments are conducted following the same setting of the first data split and features as Section 5.3. We adopt multivariate ridge regression presented in the first chapter as the comparative multi-output regression method here. PCP percentage for correctly estimated body parts are 59.03%, which are much lower than our structural learning based models, and slightly better than single-output support vector regression 58.33%. In the light of this, we can conclude that for human pose estimation, simple multi-variate output regression model

Table 5.2: PCP with different size of randomly selected training sets for VOC 2007 (including 91 testing images), where SoD denotes the size of training database.

SoD	50	100	200	400
SVR	26.32%	30.82%	31.98%	37.24%
SSVR	28.57%	44.42%	53.41%	61.44%
LSSVR	28.57%	30.15%	33.32%	40.06%

Table 5.3: PCP with different size of balanced training dataset (that is, we select equal number of images for each of five pose categories), where the testing database is the same as Table 5.1.

SoD	50	100	200	400
SVR	28.14%	29.22%	32.77%	39.21%
SSVR	47.91%	51.32%	60.11%	66.51%
LSSVR	45.27%	49.85%	60.03%	68.74%

performs worse due to missing correlated information between output entries.

5.3.7 Discussions

Firstly, it is evident from our results that discriminative methods can process test images much more efficiently. They are thus more suitable for online/real-time application, even when training database is small. Secondly, when using less training images, our methods could also achieve good results, i.e. the performance of our structured discriminative methods degrade gracefully when the training dataset size decreases. Thirdly, compared to a standard discriminative method SVR, structured techniques lead to superior results, which demonstrate the importance of introducing structured learning method to 2D human pose estimation. Finally, we could benefit from adding hidden variables into Structural Support Vector Regression when the training dataset is large enough and balanced.

5.4 Summary

This chapter has investigated three model-free discriminative methods for 2D human upper-body pose estimation. As seen from the results presented in the last section, our method could solve the problem effectively and more efficiently than previous works. Compared to generative model-based method, our techniques could not only achieve good performance but also high-efficiency owing to the nature of discriminative methods. Additionally, more benefits could be achieved by capturing the correlations between output variables using structured discriminative methods. We

also discover that compared to Latent Structural Support Vector Regression, Structural Support Vector Regression perform well given less training data and when the training data is unbalanced. Otherwise, LSSVR is preferred.

Chapter 6

Conclusion and Future Work

This thesis has set out to explore the possibility of more informative and less ambiguous visual representation and to capture the latent dependent relation in feature and label space in regression frameworks for solving a branch of computer vision problems. In particular, the thesis is geared towards solving the computer vision problems from two aspects: (1) large variation of feature representation caused by intrinsic and extrinsic conditions changes and (2) sparse and imbalanced data distribution. Owing to mining the latent dependency in feature and label space, the suffering of both challenges can be mitigated. Specifically,

- In Chapter 3, we employed a multi-output regression learning framework exploited for crowd counting to mine the latent correlation in feature space between spatially localised regions. Moreover, it can also seek support from neighbouring cells even when specific cell having no samples. In other words, such a framework can address data sparsity problem by capturing the latent dependent relation of person count in spatially-localised regions.

- In Chapter 4, a novel attribute (namely cumulative attribute) designed for regression problems is introduced by tackling with both challenges of feature variation and sparse and imbalanced data jointly. The key concept of our proposed attribute is to exploit the representation in attribute space with capturing the cumulative dependent nature of scalar-valued labels in regression.

- In Chapter 5, for learning the latent dependent relation between every entry of output, structured output learning is adopted for estimating human pose, which is highly corre-

lated in label space. Different from cumulative attribute in Chapter 4 to learn the latent dependent relation in label space implicitly, structural output learning models can explicitly discover the latent dependent relation in label space.

6.1 Future Work

Following the main storyline of this thesis about address both challenges of feature variation and sparse & imbalanced data, we can continue our work from the following directions.

6.1.1 Latent Dependency Mining via Multi-Output and Structural Learning

This thesis has presented regression-based frameworks to learn a discriminative mapping between input and output elements either by multi-output learning or structural learning. Specifically, a new robust counting-by-regression framework has been proposed in Chapter 3 for mining features and sharing correlated information between neighbouring localised regions via multi-output regression learning. Subsequently, Chapter 5 has described an approach for learning the mapping both between input and output and also between each elements of multi-dimensional outputs for human pose estimation. From the extensive experiments in Chapters 3 and 5, we can conclude that sharing information between input and output entries via multi-output and structural learning can significant improve the performance than the competitors missing such a vital information.

For future work on mining latent dependency from the structures of feature representation, there are two primary areas can be extended for crowd density estimation and human pose estimation.

- For crowd density estimation, based on the existing matrix-to-matrix mapping, we hope to construct the more informative tensor-based representation such as spatial-temporal feature to construct tensor-to-matrix mapping models to benefit from the latent structures of feature representation via tensor learning.

- Multilinear learning for tensor data in the application of human pose estimation [66] can be exploited in structural learning frameworks. In other words, we can further extend the scalar-valued output tensor learning into a structural form, which can capture the important latent dependency in multi-output labels.

6.1.2 Attribute Learning for Regression

Apart from information mining via multi-output and structural learning, this thesis has also demonstrated the importance and effectiveness of introducing attribute concept into regression techniques. In particular, unlike existing regression techniques that learn the direct mapping between low-level imagery features and scalar-valued output (i.e. age or person count), a novel cumulative attribute space is proposed in Chapter 4 to improve the model performance especially when only sparse and imbalanced data is available.

For furture work on constructing more informative attribute representation, the following directions could be considered.

- For facial age estimation, face expression of faces will lead to poor performance of age estimation due to the changes of shape and texture of faces. In the light of this, a robust model can be developed for removing the outliers of abnormal expression.

- For attribute learning, an improved cumulative attribute can be constructed with adding weighted latent information defining the difference between the output scalar and attribute position.

- Apart from the simple linear regression in inference adopted in Chapter 4, we can also introduce a novel inference method, which can take group structured sparsity into the consideration.

Bibliography

[1] H. Abdi. Partial least square regression (PLS regression). *Encyclopedia of Measurement and Statistics*, pages 740–744, 2007.

[2] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Proceedings of the IEEE Conference on Computer vision and pattern recognition*, pages 882–888, 2004.

[3] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041, 2006.

[4] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. In *Proceedings of the European Conference on Computer Vision*, pages 1–24, 2008.

[5] S. An, W. Liu, and S. Venkatesh. Face recognition using kernel ridge regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[6] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: people detection and articulated pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1021, 2009.

[7] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 41–48, 2006.

[8] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[9] Y. Benabbas, N. Ihaddadene, T. Yahiaoui, T. Urruty, and C. Djeraba. Spatio-temporal optical flow analysis for people counting. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 212–217, 2010.

[10] R. Benenson, M. Mathias, R. Timofte, and L.V. Gool. Pedestrian detection at 100 frames per second. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pages 2903–2910, 2012.

[11] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2007.

[12] M.B. Blaschko and C.H. Lampert. Learning to localize objects with structured output regression. In *Proceedings of the European Conference on Computer Vision*, pages 2–15, 2008.

[13] L. Bo and C. Sminchisescu. Structured output-associative regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2410. IEEE, 2009.

[14] L. Bo and C. Sminchisescu. Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, 87(1-2):28–52, 2010.

[15] P.V.K. Borges. Pedestrian detection based on blob motion statistics. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(2):224–235, 2013.

[16] G.J. Brostow and R. Cipolla. Unsupervised Bayesian detection of independent motion in crowds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 594–601, 2006.

[17] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):679–698, 1986.

[18] A.B. Chan and D. Dong. Generalized Gaussian process models. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pages 2681–2688, 2011.

[19] A.B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008.

[20] A.B. Chan, M. Morrow, and N. Vasconcelos. Analysis of crowded scenes using holistic properties. In *Proceedings of the IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, pages 101–108, 2009.

[21] A.B. Chan and N. Vasconcelos. Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5):909–926, 2008.

[22] A.B. Chan and N. Vasconcelos. Bayesian Poisson regression for crowd counting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 545–551. IEEE, 2009.

[23] A.B. Chan and N. Vasconcelos. Counting people with low-level features and Bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177, 2012.

[24] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. A ranking approach for human ages estimation based on face images. In *Proceedings of the International Conference on Pattern Recognition*, pages 3396–3399, 2010.

[25] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 585–592, 2011.

[26] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *Proceedings of the International Conference on Computer Vision*, pages 914–921, 2011.

[27] W.-L. Chao, J.-Z. Liu, and J.-J. Ding. Facial age estimation based on label-sensitive learning and age-oriented regression. *Pattern Recognition*, 46(3):628–641, 2013.

[28] K. Chen, S. Gong, and T. Xiang. Human pose estimation using structural support vector machines. In *Proceedings of the IEEE International Conference on Computer Vision Workshop*, pages 846– 851, 2011.

[29] K. Chen, S. Gong, T. Xiang, and C. C. Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474, 2013.

[30] K. Chen, C.C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *Proceedings of the British Machine Vision Conference*, pages 21.1–21.11, 2012.

[31] J. Cheng, J. Yang, Y. Zhou, and Y. Cui. Flexible background mixture models for foreground segmentation. *Image and Vision Computing*, 24(5):473–482, May 2006.

[32] S.Y. Cho, T.W.S. Chow, and C.T. Leung. A neural-based crowd estimation by hybrid global learning algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 29(4):535–541, 1999.

[33] Y. Cong, H. Gong, S.C. Zhu, and Y. Tang. Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pages 1093–1100, 2009.

[34] D. Conte, P. Foggia, G. Percannella, and M. Vento. A method based on the indirect approach for counting people in crowded scenes. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 111–118. IEEE, 2010.

[35] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[36] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active shape models their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[37] A. Criminisi, J. Shotton, and E. Konukoglu. Decision forest for classification, regression, density estimation, manifold learning and semi-supervised learning. Technical Report MSR-TR-2011-114, Microsoft Research, 2011.

[38] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, 2005.

[39] G. David. Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, 2:1150–1157, 1999.

[40] A.C. Davies, J.H. Yin, and S.A. Velastin. Crowd monitoring using image processing. *Electronics & Communication Engineering Journal*, 7(1):37–47, 1995.

[41] K. De Brabanter, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, 22(1):110–120, 2011.

[42] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):743–761, 2011.

[43] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami. Fast crowd segmentation using shape indexing. In *Proceedings of the IEEE International Conference on Computer Vision*, 2007.

[44] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2009.

[45] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Ferrari. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *International Journal of Computer Vision*, 99(2):190–214, 2012.

[46] N. Fan. Learning nonlinear distance functions using neural network for regression with application to robust human age estimation. In *Proceedings of the International Conference on Computer Vision*, pages 249–254, 2011.

[47] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2009.

[48] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1), 2005.

[49] V. Ferrari, M. Marin-Jiminez, and A. Zisserman. Progressive search space reduction for human pose estimation. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[50] R. Fooprateepsiri and W. Kurutach. Face verification base-on Hausdorff-shape context. In *Proceedings of the Asia Conference on Informatics in Control, Automation and Robotics*, pages 240–244, 2009.

[51] Y. Fu, G. Guo, and T.S. Huang. Age synthesis and estimation via faces: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

[52] Y. Fu, T.M. Hospedales, T. Xiang, and S. Gong. Attribute learning for understanding unstructured social activity. In *Proceedings of the European Conference on Computer Vision*, pages 530–543, 2012.

[53] Y. Fu and T.S. Huang. Human age estimation with regression on discriminative aging manifold. *IEEE Transactions on Multimedia*, 10(4):578–584, 2008.

[54] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2188–2202, 2011.

[55] W. Ge and R. Collins. Crowd detection with a multiview sampler. In *Proceedings of the European Conference on Computer Vision*, pages 324–337, 2010.

[56] W. Ge and R.T. Collins. Marked point processes for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2913–2920, 2009.

[57] P. Geladi and B.R. Kowalski. Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17, 1986.

[58] X. Geng, K. Smith-Miles, and Z.-H. Zhou. Facial age estimation by learning from label distributions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2401–2412, 2010.

[59] X. Geng, K. Smith-Miles, Z.-H. Zhou, and L. Wang. Face image modeling by multilinear subspace analysis with missing values. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 41(3):881– 892, 2011.

[60] X. Geng, Z.-H. Zhou, and K. Smith-Miles. Automatic age estimation based on facial aging patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2234–2240, 2007.

[61] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proceedings of the International Conference on Computer Vision*, pages 415–422, 2011.

[62] D. Gowsikhaa, S. Abirami, and R. Baskaran. Automated human behavior analysis from surveillance videos: a survey. *Artificial Intelligence Review*, pages 1–19, 2012.

[63] G. Guo, Y. Fu, C.R. Dyer, and T.S. Huang. Head pose estimation: classification or regression? In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4, 2008.

[64] G. Guo, Y. Fu, T.S. Huang, and C.R. Dyer. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[65] G. Guo, G. Mu, Y. Fu, and T.S. Huang. Human age estimation using bio-inspired features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 112–119, 2009.

[66] W. Guo, I. Kotsia, and I. Patras. Tensor learning for regression. *IEEE Transactions on Image Processing*, 21(2):816–827, 2012.

[67] Y. Haitovsky. On multivariate ridge regression. *Biometrika*, 74(3):563–570, 1987.

[68] R.M. Haralick, K. Shanmugam, and I.H. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics*, 3(6):610–621, 1973.

[69] J. Heikki and O. Silvn. A real-time system for monitoring of cyclists and pedestrians. In *Proceedings of the Second IEEE Workshop on Visual Surveillance*, pages 74–81, 1999.

[70] D. Helbing, I.J. Farkas, P. Molnar, and T. Vicsek. *Simulation of pedestrian crowds in normal and evacuation situations*. 2002.

[71] D. Helbing, A. Johansson, and Eidgenössische Technische Hochschule. *Pedestrian, crowd and evacuation dynamics*. Encyclopedia of Complexity and Systems Science, 2009.

[72] A.E. Hoerl and R.W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, pages 55–67, 1970.

[73] W. Hu, T. Tan, L. Wang, and S. Maybank. A survey on visual sureillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Review*, 34(3):334–352, 2004.

[74] G. Hua, M.-H. Yang, and Ying Wu. Learning to estimate human pose with data driven belief propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 747–754, 2005.

[75] Y. Huang, D. Xu, and F. Nie. Patch distribution compatible semisupervised dimension reduction for face and human gait recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(3):479–488, 2012.

[76] C. Ionescu, L. Bo, and C. Sminchisescu. Structural svm for visual localization and continuous state estimation. *Proceedings of IEEE International Conference on Computer Vision*, 2009.

[77] R. Jafri and H.R. Arabnia. A survey of face recognition techniques. *Journal of Information Processing Systems*, 5(2):41–68, 2009.

[78] T. Jan. Neural network based threat assessment for automated visual surveillance. *Proceedings of the International Joint Conference on Neural Networks*, 2:1098–7576, 2004.

[79] T. Joachims, T. Finley, and Chun-Nam Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009.

[80] S. Johnson and M. Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. *Proceedings of the IEEE International Conference on Computer Vision*, 2009.

[81] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. *Proceedings of the British Machine Vision Conference*, 2010.

[82] C. Jordan and B. Taskar. Adaptive pose priors for pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 422–429, 2010.

[83] A. Kanaujia, C. Sminchisescu, and D.N. Metaxas. Semi-supervised hierarchical models for 3D human pose reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[84] W. Kang and F. Deng. Research on intelligent visual surveillance for public security. *Proceedings of the IEEE/ACIS International Conference on Computer and Information Science*, pages 824–829, 2007.

[85] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the International Conference on Machine Learning*, pages 1–8, 2011.

[86] T. Ko. A survey on behavior analysis in video surveillance for homeland secrity applications. *IEEE Applied Imagery Pattern Recognition Workshop*, pages 1–8, 2008.

[87] D. Kong, D. Gray, and H. Tao. Counting pedestrians in crowds using viewpoint invariant training. In *Proceedings of the British Machine Vision Conference*, pages 1–10, 2005.

[88] D. Kong, D. Gray, and H. Tao. A viewpoint invariant approach for crowd counting. In *Proceedings of the International Conference on Pattern Recognition*, volume 3, pages 1187–1190, 2006.

[89] N. Krahnstoever and P.R.S. Mendonca. Bayesian autocalibration for surveillance. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2, pages 1858–1865. IEEE, 2005.

[90] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1446–1453, 2009.

[91] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

[92] C.H. Lampert. *Kernel methods in computer vision*. Now Publishers Inc, 2009.

[93] C.H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[94] A. Lanitis, C. Draganova, and C. Christodoulou. Comparing different classifiers for automatic age estimation. *IEEE Transactions onSystems, Man, and Cybernetics, Part B: Cybernetics*, 34(1):621–628, 2004.

[95] B. Laxton. Moncular human pose estimation. *University of California, San Diego, Research Examination for the MS degree*, pages 1–16, 2007.

[96] R. Layne, T. Hospedales, and S. Gong. Person re-identification by attributes. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2012.

[97] M.W. Lee and I. Cohen. Human upper body pose estimation in static images. In *Proceedings of the European Conference on Computer Vision*, pages 126–138, 2004.

[98] M.W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Proceedings of the IEEE Conference on Computer vision and pattern recognition*, pages 334–341, 2004.

[99] M.W. Lee and R. Nevatia. Body part detection for human pose estimation and tracking. In *Proceedings of the IEEE Workshop on Motion and Video Computing*, pages 23–23, 2007.

[100] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, volume 1, pages 878–885, 2005.

[101] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 1324–1332, 2010.

[102] V. Lepetit and P. Fua. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends in Computer Graphics and Vision*, 81:1–89, 2005.

[103] J. Li, L. Huang, and C. Liu. CASIA pedestrian counting dataset: `http://cpcd.vdb.csdb.cn/page/showItem.vpage?id=automation.dataFile/1`.

[104] J. Li, L. Huang, and C. Liu. Robust people counting in video surveillance: Dataset and system. In *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance*, pages 54–59. IEEE, 2011.

[105] M. Li, Z. Zhang, K. Huang, and T. Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Proceedings of the International Conference on Pattern Recognition*, pages 1–4, 2008.

[106] A. Liaw and M. Wiener. Classification and regression by random forest. *R News*, 2(3):18–22, 2002.

[107] S.F. Lin, J.Y. Chen, and H.X. Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, 31(6):645–654, 2001.

[108] T.Y. Lin, Y.Y. Lin, M.F. Weng, Y.C. Wang, Y.F. Hsu, and H.Y.M. Liao. Cross camera people counting with perspective estimation and occlusion handling. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pages 1–6, 2011.

[109] J. Liu, R.T. Collins, and Y. Liu. Surveillance camera autocalibration based on pedestrian height distributions. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2011.

[110] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344. IEEE, 2011.

[111] Y. Long. Human age estimation by metric learning for regression problems. In *Proceedings of the International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 455–465, 2009.

[112] C. C. Loy, T. Xiang, and S. Gong. Time-delayed correlation analysis for multi-camera activity understanding. *International Journal of Computer Vision*, 90(1):106–129, 2010.

[113] C. C. Loy, T. Xiang, and S. Gong. Salient motion detection in crowded scenes. In *Proceedings of the International Symposium on Communications, Control and Signal Processing*, pages 1–4, 2012.

[114] C.C. Loy, K. Chen, S. Gong, and T. Xiang. *Crowd Counting and Profiling: Methodology and Evaluation*. 2013.

[115] K. Luu, K. Ricanek, T.D. Bui, and C.Y. Suen. Age estimation using active appearance models and support vector machine regression. In *Proceedings of the IEEE International Conference on Biometrics: Theory, Applications and Systems*, pages 314–318, 2009.

[116] R. Ma, L. Li, W. Huang, and Q. Tian. On pixel count based crowd density estimation for visual surveillance. In *Proceedings of the IEEE Conference on Cybernetics and Intelligent Systems*, volume 1, pages 170–173, 2004.

[117] W. Ma, L. Huang, and C. Liu. Crowd density analysis using co-occurrence texture features. In *Proceedings of the International Conference on Computer Sciences and Convergence Information Technology*, pages 170–175, 2010.

[118] A. Manzanera and J.C. Richefeu. A new motion detection algorithm based on [Sigma]-[Delta] background estimation. *Pattern Recognition Letters*, 28(3):320–328, 2007.

[119] A.N. Marana, L. F. Costa, R.A. Lotufo, and S.A. Velastin. Estimating crowd density with minkowski fractal dimension. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 6, pages 3521–3524, 1999.

[120] A.N. Marana, L.F. Costa, R.A. Lotufo, and S.A. Velastin. On the efficacy of texture analysis for crowd monitoring. In *Proceedings of the International Symposium on Computer Graphics, Image Processing, and Vision*, pages 354–361, 1998.

[121] A.N. Marana, S.A. Velastin, L.F. Costa, and R.A. Lotufo. Estimation of crowd density using image processing. In *Proceedings of the Image Processing for Security Applications*, pages 1–11, 1997.

[122] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behaviour detection using social force model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 935–942, 2009.

[123] T. B. Moeslund and E. Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 81:231–268, 2000.

[124] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126, 2006.

[125] A. Montillo and H. Ling. Age regression from faces using random forests. In *IEEE International Conference on Image processing*, pages 2437–2440, 2009.

[126] G. Mori and J. Malik. Estimating human body configurations using shape context matching. In *Proceedings of the European Conference on Computer Vision*, pages 666–680, 2002.

[127] E. Murphy-Chutorian and M.M. Trivedi. Head pose estimation in computer vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[128] R. Navaratnam, A. W. Fitzgibbon, and R. Cipolla. The joint manifold model for semi-supervised multi-valued regression. In *Proceedings of the International Conference on Computer Vision*, pages 1 – 8, 2007.

[129] R. Nevatia and C.-W. Chu. Body pose estimation and gesture recognition for human-computer interaction system. *University of Southern California, Doctoral Dissertation*, 2008.

[130] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *Proceedings of the ACM International Conference on Multimedia*, pages 85–94, 2009.

[131] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[132] R. Okada and S. Soatto. Relevant feature selection for human pose estimation and localization in cluttered images. In *Proceedings of the European Conference on Computer Vision*, pages 434–445, 2008.

[133] M. Pätzold, R.H. Evangelio, and T. Sikora. Counting people in crowded environments by fusion of shape and motion information. In *Proceedings of the IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 157–164, 2010.

[134] K. Peng, L. Chen, and S. Ruan. A novel scheme of face verification using active appearance models. In *Proceedings of the International Conference on Advanced Video and Signal-Based Surveillance*, pages 247–252, 2005.

[135] T. Pfister, J. Charles, M. Everingham, and A. Zisserman. Automatic and efficient long term arm and hand tracking for continuous sign language TV broadcasts. In *Proceedings of the British Machine Vision Conference*, pages 1–11, 2012.

[136] N. Pourdamghani, H.R. Rabiee, F. Faghri, and M. H. Rohban. Graph based semi-supervised human pose estimation: When the output space comes to help. *Pattern Recognition Letters*, 33(12):1529–1535, 2012.

[137] V. Rabaud and S. Belongie. Counting crowded moving objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 705–711, 2006.

[138] D. Ramanan. Learning to parse images of articulated objects. In *Proceedings of the Advanced Neural Information Processing Systems*, pages 1–8, 2006.

[139] C. E. Rasmussen and C. K. I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006.

[140] X. Ren, A.C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–831, 2005.

[141] K. Rifcanek and T. Tesafaye. Morph: a longitudinal image database of normal adult age-progression. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 341–345, 2006.

[142] M. Rodriguez, I. Laptev, J. Sivic, and J.Y. Audibert. Density-aware person detection and tracking in crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2423–2430, 2011.

[143] D. Russell and S. Gong. Minimum cuts of a time-varying background. In *Proceedings of the British Machine Vision Conference*, pages 809–818, 2006.

[144] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *Proceedings of the Digital Image Computing: Techniques and Applications*, pages 81–88, 2009.

[145] P. Sabzmeydani and G. Mori. Detecting pedestrians by learning shapelet features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[146] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the International Conference on Machine Learning*, pages 515–521, 1998.

[147] C. Shan, S. Gong, and P.W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803–816, 2009.

[148] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[149] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[150] V.K. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 314–327, 2010.

[151] A.J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[152] M. Solnon, S. Arlot, and F. Bach. Multi-task regression using minimal penalties. *Journal of Machine Learning Research*, 13:2773–2812, 2012.

[153] J. Spehr, S. Winkelbach, and F.M. Wahl. Hierarchical pose estimation for human gait analysis. *Comput. Methods Prog. Biomed.*, 106(2):104–113, 2012.

[154] C. Stauffer and W. E. L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[155] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.

[156] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3394–3401, 2012.

[157] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1616–1623, 2012.

[158] A. Sundaresan and R. Chellappa. Multi-camera tracking of articulated human motion

using motion and shape cues. In *Proceedings of the Asian conference on Computer Vision*, pages 131–140, 2006.

[159] J. A. K. Suykens, J. De Brabanter, B. De Moor, and J. Vandewalle. Automatic relevance determination for least squares support vector machine regression. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2416–2421, 2001.

[160] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3):293–300, 1999.

[161] Y. Tian, L. Brown, A. Hampapur, M. Lu, A. Senior, and C. Shu. Ibm smart surveillance system (s3): event based video surveillance system with an open and extensible framework. *Machine Vision and Applications*, 19(5):315–327, 2008.

[162] Y. Tian, L. Sigal, H. Badino, F. De la Torre Frade, and Y. liu. Latent Gaussian mixture regression for human pose estimation. In *Proceedings of the Asian Conference on Computer Vision*, pages 1–12, 2010.

[163] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *Proceedings of the International Conference on Machine Learning*, 2004.

[164] J. Tu, Y. Fu, and T.S. Huang. Locating nose-tips and estimating head poses in images by tensorposes. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(1):90–102, 2009.

[165] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Unified crowd segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 691–704, 2008.

[166] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

[167] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[168] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York Inc, 2000.

[169] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[170] P. Viola, M.J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161, 2005.

[171] C.-C. Wang, Y.-C. Su, C.-T. Hsu, C.-W. Lin, and H. Y. M. Liao. Bayesian age estimation on face images. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 282–285, 2009.

[172] L. Wang, W. Hu, and T. Tan. Recent developments in motion analysis. *Pattern Recognition*, 36(3):585–601, 2003.

[173] M. Wang, W. Li, and X. Wang. Transferring a generic pedestrian detector towards specific scenes. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pages 3274–3281, 2012.

[174] M. Wang and X. Wang. Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3401–3408. IEEE, 2011.

[175] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 710–724, 2008.

[176] M. Welling. Support vector regression. Technical report, Department of Computer Science, University of Toronto, 2004.

[177] J. Weston, B. Schoelkopf, O. Bousquet, T. Mann, and W. S. Noble. Joint kernel maps. *Proceedings of the International Conference on Computer Vision*, pages 67–83, 2004.

[178] B. Wu and R. Nevatia. Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Proceedings of the International Conference on Computer Vision*, volume 1, pages 90–97, 2005.

[179] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *International Journal of Computer Vision*, 75(2):247–266, 2007.

[180] X. Wu, G. Liang, K.K. Lee, and Y. Xu. Crowd density estimation using texture analysis and learning. In *Proceedings of the IEEE International Conference on Robotics and Biomimetics*, pages 214–219, 2006.

[181] S. Yan, H. Wang, T.S. Huang, Q. Yang, and X. Tang. Ranking with uncertain labels. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 96–99, 2007.

[182] S. Yan, H. Wang, X. Tang, and T.S. Huang. Learning auto-structured regressor from uncertain nonnegative labels. In *Proceedings of the International Conference on Computer Vision*, pages 1–8, 2007.

[183] S. Yan, X. Zhou, M. Liu, M. Hasegawa-Johnson, and T.S. Huang. Regression from patch-kernel. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[184] D.B. Yang, H.H. González-Baños, and L.J. Guibas. Counting people in crowds with a real-time network of simple image sensors. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 122–129, 2003.

[185] P. Yang, L. Zhong, and D. Metaxas. Ranking model for facial age estimation. In *Proceedings of the International Conference on Pattern Recognition*, pages 3404–3407, 2010.

[186] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392, 2011.

[187] A. Yao, J. Gall, G. Fanelli, and L. V. Gool. Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11, 2011.

[188] O. Yeniay and A. Goktas. A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31:99–111, 2002.

[189] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Journal of Computing Surveys*, 38(4):1–45, 2006.

[190] C.-N. Yu and T. Joachims. Learning structural SVMs with latent variables. *Proceedings of the International Conference on Machine Learning*, pages 1–8, 2009.

[191] W. Yu. Face recognition using constrained active appearance model. In *Proceedings of the International Symposium on Intelligent Information Technology Application Workshops*, pages 348–351, 2009.

[192] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915–936, 2003.

[193] J. Zhang, B. Tan, F. Sha, and L. He. Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1037–1046, 2011.

[194] Y. Zhang and D. Yeung. Multi-task warped gaussian process for personalized age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2622–2629, 2010.

[195] T. Zhao, R. Nevatia, and B. Wu. Segmentation and tracking of multiple humans in crowded environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1198–1211, 2008.

[196] B. Zhou, X. Wang, and X. Tang. Random field topic model for semantic region analysis in crowded scenes from tracklets. In *Proceedings of the IEEE Conference Computer Vision and Pattern Recognition*, pages 3441–3448, 2011.